TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

PHẠM THỊ ÁNH PHÁT

PHƯƠNG PHÁP XÁO TRỘN DỮ LIỆU CHO ĐẶC TRƯNG TỪ TRONG GIẢI THÍCH TÁC VỤ GÁN NHÃN NGỮ NGHĨA Y SINH

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN CNTT
CHƯƠNG TRÌNH CHẤT LƯỢNG CAO

TP.HCM, NĂM 2024

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

PHẠM THỊ ÁNH PHÁT – 20127680

PHƯƠNG PHÁP XÁO TRỘN DỮ LIỆU CHO ĐẶC TRƯNG TỪ TRONG GIẢI THÍCH TÁC VỤ GÁN NHÃN NGỮ NGHĨA Y SINH

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN CNTT
CHƯƠNG TRÌNH CHẤT LƯỢNG CAO

GIẢNG VIÊN HƯỚNG DẪN
ThS. TUẤN NGUYÊN HOÀI ĐỨC

TP. HCM, NĂM 2024

Lời cảm ơn

Lời đầu tiên, em xin bày tỏ lòng biết ơn sâu sắc đến Thạc sĩ Tuấn Nguyên Hoài Đức – người đã tận tình hướng dẫn, hỗ trợ và tạo mọi điều kiện thuận lợi nhất cho em trong suốt thời gian thực hiện đề tài. Thầy đã tận tâm chỉ bảo từ cách đọc, cách bắt đầu nghiên cứu một bài báo đến lúc chỉnh sửa hoàn thiện quyển báo cáo luận văn. Em thật sự rất biết ơn và trân quý những kiến thức và kinh nghiệm mà Thầy đã truyền đạt; chắc chắn đó sẽ là hành trang quý báu cho con đường sự nghiệp của em sau này.

Em cũng xin chân thành cảm ơn các Thầy, Cô trong khoa Công nghệ Thông tin – Trường Đại học Khoa học Tự nhiên đã giảng dạy, truyền đạt kiến thức và tạo môi trường học tập đầy đổi mới, sáng tạo để khai thác tối đa thế mạnh của sinh viên.

Em cũng xin gửi lời cảm ơn đến ban giám hiệu và các phòng ban chức năng của trường đã tạo điều kiện thuận lợi, xây dựng một chương trình học tập luôn đổi mới để sinh viên có thể bắt kịp với nhịp độ của thời đại. Đồng thời, em xin chân thành cảm ơn gia đình, bạn bè, người thân đã luôn bên cạnh động viên, chia sẻ, là nguồn động lực rất lớn để em luôn phấn đấu, hoàn thành tốt nhiệm vụ của mình.

Bên cạnh đó, em còn những thiếu sót trong quá trình hoàn thành khóa luận, em rất mong nhận được những góp ý, đánh giá của các Thầy, Cô giáo để em có thể hoàn thiện hơn. Cuối cùng, em xin kính chúc các Thầy, Cô trong ban lãnh đạo, trong khoa có thật nhiều sức khỏe và luôn đạt được thành công trong sự nghiệp trồng người của mình.

Trân trọng.

TP.Hồ Chí Minh, tháng 7 năm 2024

Sinh viên thực hiện

Phạm Thị Ánh Phát

Mục lục

Danh sách các hình

Danh sách các bảng

Tóm tắt

Các mô hình học sâu (Deep Learning) hiện nay đã đạt được những thành tựu ấn tượng trong rất nhiều lĩnh vực như thị giác máy tính, nhận diện thực thể, xử lý ngôn ngữ tự nhiên. Nổi bật nhất hiện nay là sự ra đời của ChatGPT-4 đang dần thống lĩnh hệ thống chatbot. Tuy nhiên, sự phát triển vượt bậc của các hệ thống Trí tuệ Nhân tạo (AI) hiện nay vẫn đặt ra mối lo ngại về mức độ an toàn và khả năng giải thích của chúng bởi bản chất phức tạp và khó hiểu từ các mô hình.

Để giải quyết vấn đề này, Trí tuệ Nhân tạo Khả diễn giải (EXplainable AI - XAI) đã trở thành một lĩnh vực nghiên cứu quan trọng. Nhờ vào XAI mà các mô hình học sâu có thể được diễn giải và dễ hiểu hơn với người dùng, tạo độ tin cậy và thúc đẩy sự chấp nhận vào các quyết định của mô hình. Đặc biệt khi các hệ thống AI được ứng dụng vào trong các lĩnh vực mang tính quyết định như tài chính, pháp lý và quan trọng hơn cả là lĩnh vực y học. Việc giúp máy tính có thể hiểu và khai thác tri thức trong Y học để đưa ra các quyết định chính xác là rất cần thiết, đó cũng chính là mục tiêu của tác vụ Gán nhãn Ngữ nghĩa trong văn bản Y sinh. Tuy nhiên, để người dùng có thể đặt niềm tin vào các quyết định của hệ thống, các nhà phát triển cần cung cấp cho người dùng cách mà mô hình tính toán và hình thành các dự đoán.

Chính vì thế mà khóa luận đã đề xuất một phương pháp giải thích dựa trên đặc trưng từ của mô hình ngôn ngữ trên tác vụ Gán nhãn Ngữ nghĩa trong văn bản Y sinh. Mục đích của phương pháp này là sử dụng các kỹ thuật xáo trộn dữ liệu để đánh giá tầm quan trọng của đặc trưng từ một cách khách quan và chính xác nhất. Bên cạnh đó, khóa luận còn thực hiện việc khám phá không gian ẩn (vector biểu diễn) của mô hình tác vụ trong việc hình thành kết quả dự đoán của mô hình.

Chương 1: GIỚI THIỆU ĐỀ TÀI

1.1 Lí do chọn đề tài

Học sâu (Deep Learning) đang là tâm điểm trong nghiên cứu và ứng dụng Trí tuệ nhân tạo (AI) bởi tính hiệu quả mà chúng mang lại trên rất nhiều lĩnh vực. Điển hình như sự xuất hiện của các hệ thống chatbot như ChatGPT, Gemini đang làm nổi lên xu hướng thần thánh hóa các mô hình Trí tuệ nhân tạo. Tuy nhiên, tính phức tạp và thiếu minh bạch của các mô hình AI hiện đại, đặc biệt trong Xử lý ngôn ngữ tự nhiên (NLP), đã gây ra nhiều lo ngại về độ tin cậy và tính công bằng. Người dùng gặp rất nhiều khó khăn trong việc hiểu rõ cách mà AI đưa ra kết quả, dẫn đến thiếu tin tưởng và khó khăn trong việc chấp nhận các giải pháp AI.

Để giải quyết vấn đề này, Trí tuệ Nhân tạo Khả diễn giải (EXplainable AI - XAI) xuất hiện như một giải pháp tiên tiến, cung cấp cái nhìn rõ ràng và minh bạch về quá trình ra quyết định của AI. XAI không chỉ giúp người dùng hiểu và tin tưởng vào các hệ thống AI mà còn hỗ trợ các nhà nghiên cứu và phát triển trong việc cải tiến mô hình một cách hiệu quả hơn. Ngoài ra, trong các ứng dụng quan trọng như quốc phòng, chăm sóc sức khỏe, tài chính và pháp luật, khả năng giải thích của AI là vô cùng quan trọng để đảm bảo tính công bằng và tránh các sai sót nghiêm trọng.

Nhận thức được tầm quan trọng và tiềm năng của XAI trong việc nâng cao độ tin cậy và tính minh bạch trong các hệ thống Trí tuệ Nhân tạo. Đặc biệt khi được áp dụng trong lĩnh vực Y học, nơi mà máy tính có thể giúp con người đưa ra các quyết định quan trọng. Vậy làm thế nào để máy tính, không phải con người, có khả năng xử lý và hiểu được các tài liệu y sinh? Đây chính là vai trò của bài toán Gán nhãn Ngữ nghĩa (SRL) trong văn bản Y sinh, từ đó giúp máy tính hỗ trợ các bác sĩ đưa ra quyết định ảnh hưởng trực tiếp đến sức khỏe và tính mạng của con người. Do đó, việc áp dụng XAI vào trong hệ thống Xử lý Ngôn ngữ Tự nhiên, đặc biệt là tác vụ SRL đóng vai trò rất quan trọng.

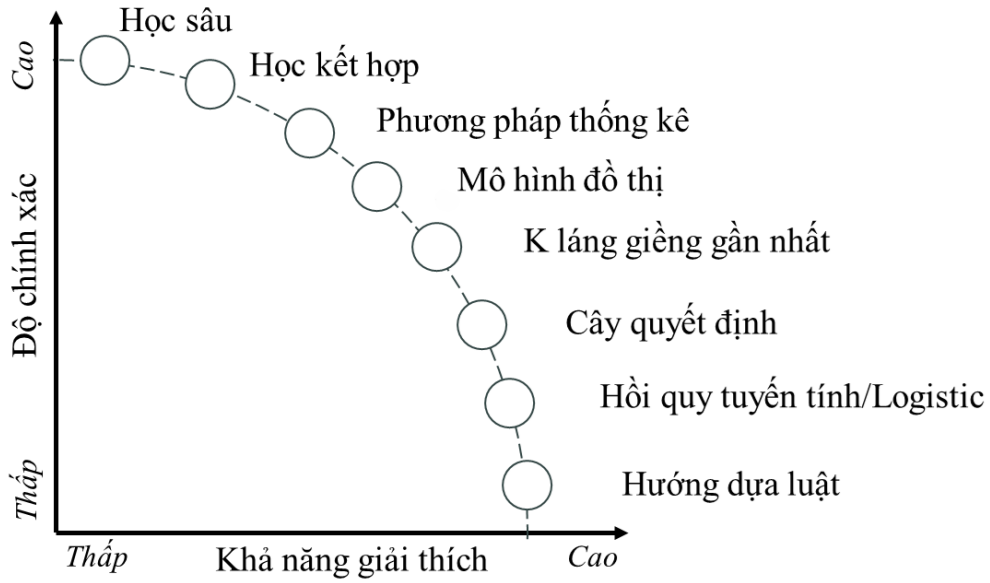
Đó cũng chính là đối tượng nghiên cứu mà khóa luận muốn hướng tới: XAI trong Xử lý Ngôn ngữ Tự nhiên (gọi tắt là X-NLP), cụ thể là cho tác vụ SRL trên tập văn bản Y Sinh. Nghiên cứu này không chỉ đóng góp về mặt học thuật mà còn có ý nghĩa thực tiễn cao, cung cấp những phân tích sâu sắc và đáng tin cậy, góp phần vào sự tiến bộ của nghiên cứu và thực hành y khoa.

1.2 Giới thiệu đề tài

Trong lĩnh vực Xử lý Ngôn ngữ Tự nhiên (NLP), các phương pháp tiếp cận đã trải qua một sự chuyển biến đáng kể. Ban đầu, các phương pháp "hộp trắng" như hệ thống dựa trên luật, cây quyết định, mô hình Markov ẩn và hồi quy logistic chiếm ưu thế nhờ tính minh bạch của chúng. Tuy nhiên, những tiến bộ gần đây đã chứng kiến sự xuất hiện của các mô hình mạng nơ-ron quy mô lớn như BERT và các biến thể của nó [11,20,58,80], vốn mang lại hiệu suất vượt trội nhưng lại thiếu tính diễn giải do cấu trúc phức tạp, được gọi là các mô hình “hộp đen”.

Sự thiếu minh bạch này đã làm dấy lên những lo ngại về thiên vị và đạo đức, có thể làm suy giảm niềm tin vào các mô hình AI [8,27,48]. Vấn đề này đặc biệt nghiêm trọng trong các lĩnh vực quan trọng như y tế, tư pháp hình sự và tài chính [62] cũng như trong các ứng dụng NLP có ảnh hưởng sâu rộng đến xã hội [22]. Những lo ngại này nhấn mạnh nhu cầu về một ngành khoa học đang phát triển, được gọi là Trí tuệ Nhân tạo Khả diễn giải cho xử lý ngôn ngữ tự nhiên (X-NLP).

Các nghiên cứu về X-NLP tập trung vào việc phát triển các mô hình cũng như các phương pháp để đạt được tính minh bạch trong NLP, đồng thời tiết lộ các nguyên nhân để con người có thể kịp thời can thiệp khi có sai sót trong kết quả từ mô hình. Điều này bao gồm việc xác định các yếu tố ảnh hưởng đến dự đoán của mô hình, có thể bao gồm các đặc trưng đầu vào, cơ chế nội bộ của mô hình hoặc tập dữ liệu huấn luyện [5]. ***Hình 1*** minh họa sự đánh đổi giữa khả năng diễn giải của mô hình và hiệu suất trong quá trình phát triển AI, với XAI đóng vai trò là cầu nối để thu hẹp khoảng cách này [24].



***Hình 1**. Minh học đánh đổi giữa độ chính xác và khả năng giải thích của mô hình.*

1.3 Thách thức bài toán và hướng giải quyết của đề tài

XAI vẫn còn là hướng nghiên cứu khá mới đối với cộng đồng nghiên cứu và đang phải đối mặt với nhiều thách thức. Thứ nhất, sự phức tạp trong khả năng giải thích rất đa dạng trong các lĩnh vực khác nhau, đòi hỏi các chiến lược giải thích đặc thù [19]. Chẳng hạn như văn bản Y sinh chứa đựng nhiều thuật ngữ chuyên ngành và ngữ cảnh đặc thù, đòi hỏi mô hình không chỉ hiểu được ngữ nghĩa mà còn phải giải thích được quá trình ra quyết định một cách chính xác vì nó ảnh hưởng trực tiếp đến việc chẩn đoán và điều trị bệnh. Hơn nữa, do tính chất đặc thù của dữ liệu y sinh, việc xây dựng bộ dữ liệu kiểm tra và thiết lập các tiêu chí đánh giá chính xác là một thách thức không nhỏ. Các mô hình XAI phải được kiểm chứng thông qua các thí nghiệm thực tiễn để đảm bảo rằng chúng không chỉ hoạt động tốt về mặt kỹ thuật mà còn mang lại giá trị thực tiễn cho các chuyên gia y tế.

Để giải quyết những thách thức này, khóa luận chỉ tập trung vào X-NLP, cụ thể là phương pháp giải thích cho tác vụ SRL trong văn bản Y sinh. Đề tài đề xuất phương

pháp nhằm định lượng chính xác sức ảnh hưởng (influence) và độ hữu ích (relevance) của từng thực từ trong dữ liệu cho dự đoán bài toán Gán nhãn Ngữ nghĩa (SRL). Phương pháp này sẽ thay thế các thực từ trong văn bản gốc bằng các từ khác có vai trò ngữ pháp tương tự nhưng ngữ nghĩa đủ khác biệt để xác định rõ vai trò của từng từ trong ngữ cảnh, giúp đánh giá tầm quan trọng của mỗi từ một cách trung thực và khách quan nhất.

Đồng thời, khóa luận cũng tập trung phân tích không gian latent của dữ liệu, xem xét liệu các vector biểu diễn của mỗi từ có phần nào phản ánh được tầm quan trọng và sức ảnh hưởng của từ đó cho tác vụ NLP đang xét hay không. Hơn nữa, vì chỉ tập trung cho văn bản Y sinh, khóa luận sẽ giải thích tác vụ SRL trên mô hình tiền huấn luyện BioBert – được thiết kế dành riêng cho dữ liệu Y sinh. Bằng cách này, đề tài không chỉ đóng góp vào việc nâng cao tính minh bạch và khả năng giải thích của các mô hình NLP mà còn đặc biệt hữu ích cho việc xử lý và phân tích văn bản Y sinh, hỗ trợ các chuyên gia y tế trong việc chẩn đoán và ra quyết định kịp thời, chính xác.

1.4 Mục tiêu đề tài

Đề tài góp phần phát triển một số phương pháp tạo ra bộ dữ liệu bị xáo trộn về mặt ngữ nghĩa ở mức từ vựng nhưng đảm bảo cấu trúc về ngữ pháp từ bộ dữ liệu ban đầu để đánh giá về hành vi của mô hình SRL. Sau đó, các bộ dữ liệu được tạo ra theo từng phương pháp xáo trộn dữ liệu được thực nghiệm trên tác vụ SRL trên văn bản Y sinh. Qua đó, khóa luận có thể đánh giá và so sánh mức độ hiệu quả của các phương pháp xáo trộn dữ liệu này trong việc mang lại lời giải thích trung thực về tầm quan trọng của các thực từ (content word) đối với dự đoán của mô hình SRL.

Khóa luận hướng tới việc đề xuất phương pháp Smart Substitution để thực hiện việc xáo trộn dữ liệu hiệu quả nhất nhằm định lượng được sức ảnh hưởng (influence) và độ hữu ích (relevance) của mỗi thực từ về mặt ngữ nghĩa một cách trung thực và khách quan nhất có thể.

Bên cạnh đó, khóa luận khám phá không gian latent của dữ liệu (tức vector biểu diễn từ) để kiểm chứng bằng thực nghiệm xem không gian latent có mã hóa tầm quan trọng của mỗi từ hay không.

Cấu trúc báo cáo như sau: **Chương 1** giới thiệu khái quát về ý nghĩa của X-NLP và mục tiêu của đề tài. **Chương 2** phác thảo các khái niệm cơ bản liên quan đến X-NLP. **Chương 3** khảo sát chi tiết các phương pháp X-NLP khác nhau để tạo ra lời giải thích dựa trên tầm quan trọng của đặc trưng. Trong khi đó, **Chương 4** và **Chương 5** liên quan đến phương pháp đề xuất của khóa luận và kết quả thực nghiệm cũng như các thảo luận liên quan. Cuối cùng, **Chương 6** nêu kết luận và các hướng phát triển tiềm năng của đề tài.

Chương 2: CƠ SỞ LÝ THUYẾT

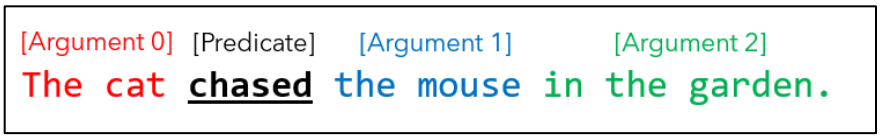
Mục tiêu cung cấp các kiến thức nền tảng về X-NLP trong tác vụ SRL, bao gồm các khía cạnh chính như Tác vụ Gán nhãn Ngữ nghĩa (SRL) (**Mục 2.1**), Tầm quan trọng của tính Khả diễn giải (**Mục 2.2**), Phạm vi giải thích (**Mục 2.3**), Chất lượng giải thích (**Mục 2.4**) và Phân loại giải thích (**Mục 2.5**).

2.1 Tác vụ Gán nhãn Ngữ nghĩa (Semantic Role Labeling - SRL)

2.1.1 Cấu trúc Đối số Vị ngữ (Predicate-Argument Structures - PAS)

Để máy tính có khả năng đọc hiểu và xử lý được lượng tri thức thông qua các văn bản Y sinh, máy tính cần nắm được cấu trúc cũng như nội dung của từng câu trong văn bản. Cấu trúc của một câu thông thường luôn bao gồm động từ chính, gọi là vị ngữ (predicate), và các đối tượng liên quan đến vị ngữ (chủ ngữ, bổ ngữ,…), gọi là các đối số (argument).

Trong cấu trúc này, vị ngữ đóng vai trò trung tâm, xoay quanh là các đối số có vai trò ngữ nghĩa cụ thể. Đây cũng chính là Cấu trúc Đối số Vị ngữ (Predicate Argument Structure - PAS). Máy tính có thể hoàn toàn hiểu được nội dung cấu trúc của câu dựa vào tri thức mà PAS cung cấp. ***Hình 2*** là một ví dụ minh họa về PAS:



Hình 2. *Ví dụ về Cấu trúc Đối số Vị ngữ (PAS) trong một câu.*

Vị ngữ "chased" đóng vai trò là sự kiện chính trong câu, xung quanh là các đối số để bổ trợ ngữ nghĩa cho vị ngữ như:

Đối số 1: Đối tượng hành động.

Đối số 2: Đối tượng bị hành động.

Đối số 3: Địa điểm hành động.

Cấu trúc PAS giúp làm rõ các thành phần trong câu liên kết với nhau và vai trò ngữ pháp của chúng. Nhờ đó, PAS là công cụ hữu ích cho hàng loạt các bài toán Xử lý Ngôn ngữ Tự nhiên (NLP), trong đó có tác vụ Gán nhãn Ngữ nghĩa (SRL). SRL là bài toán nhằm xác định vai trò ngữ nghĩa của các thành phần trong câu. Đặc biệt trong lĩnh vực Y sinh, SRL giúp phân tích và hiểu sâu hơn các mối quan hệ giữa các thực thể và sự kiện y học, từ đó hỗ trợ nghiên cứu, chẩn đoán và điều trị bệnh. SRL không chỉ giúp tự động hóa quá trình phân tích và hiểu văn bản mà còn làm giàu thêm kho dữ liệu Y sinh, hỗ trợ các hệ thống thông tin y tế, nghiên cứu y học và quyết định lâm sàng.

Cụ thể nhiệm vụ của SRL như sau:

- Xác định Vị ngữ: tìm các động từ hoặc sự kiện chính trong câu
- Xác định Đối số: tìm các thực thể hoặc cụm từ đóng vai trò bổ nghĩa cho sự kiện chính trong câu.
- Gán nhãn Vai trò Ngữ nghĩa: Gán nhãn cho từng đối số với vai trò ngữ nghĩa thích hợp như chủ thể, địa điểm, đối tượng, thời gian.

Để xây dựng một mô hình SRL hiệu quả trong lĩnh vực y học rất cần thiết để có một bộ dữ liệu huấn luyện chất lượng cao. Các nguồn dữ liệu phong phú và quy trình chú thích cẩn thận sẽ giúp đảm bảo rằng mô hình có thể nhận diện và gán nhãn chính xác các thực thể và vai trò ngữ nghĩa trong các văn bản y sinh phức tạp. Các bộ dữ liệu huấn luyện có thể được lấy từ PubMed – cơ sở dữ liệu lớn của các bài báo nghiên cứu Y sinh và khoa học đời sống, MEDLINE – cơ sở dữ liệu bao gồm các bài tóm tắt và trích dẫn của các bài báo y khoa, GENIA và BioNLP Shared Task – các bộ dữ liệu chứa các câu được chú thích với thông tin về thực thể sinh học và các mối quan hệ ngữ nghĩa.

2.2 Tầm quan trọng của tính Khả diễn giải

2.2.1 An toàn, độ tin cậy và vấn đề đạo đức trong NLP

An toàn và độ tin cậy là hai khái niệm then chốt trong các nền tảng NLP như trợ lý ảo hoặc hệ thống tư vấn. Người dùng cần tin tưởng vào khả năng xử lý của mô hình NLP, ngay cả khi đối mặt với các cấu trúc ngôn ngữ mới lạ [42]. Sự tin tưởng này đảm bảo rằng chúng sẽ không tạo ra các kết quả không phù hợp, gây hiểu lầm hoặc có hại. Việc thiếu minh bạch trong các dự đoán có thể dẫn đến khủng hoảng [13]. Do đó, việc đảm bảo an toàn tối đa giúp các nhà phát triển và người dùng tin tưởng vào hiệu quả và độ an toàn của mô hình trong các ngữ cảnh ngôn ngữ đa dạng, đồng thời nâng cao uy tín thương hiệu và tạo điều kiện cho các mối quan hệ kinh doanh lâu dài [10].

Hơn nữa, vấn đề đạo đức trong NLP rất quan trọng vì các mô hình này liên quan trực tiếp với ngôn ngữ, văn hóa và các sắc thái của con người. Các mô hình NLP hoạt động trong các bối cảnh ngôn ngữ đa dạng và có thể vô tình duy trì các định kiến hoặc thiên vị trong xã hội [8,27]. Khả năng diễn giải là cần thiết để hiểu và đảm bảo rằng các mô hình đưa ra quyết định không thiên vị, phù hợp với chuẩn mực đạo đức và các giá trị toàn cầu [48]. Điều này thúc đẩy việc sử dụng AI có trách nhiệm, tránh những tác hại tiềm ẩn và thúc đẩy sự công bằng trong xử lý ngôn ngữ.

2.2.2 Trách nhiệm giải trình trong các hệ thống NLP

Các hệ thống NLP, được sử dụng rộng rãi trên nền tảng mạng xã hội và trong các hệ thống tư vấn, ảnh hưởng đáng kể đến nhận thức và hành vi của cộng đồng. Trách nhiệm giải trình trong NLP là vô cùng quan trọng, bao gồm việc hiểu rõ quy trình ra quyết định, cho phép nhà phát triển phát hiện và kịp thời khắc phục các lỗ hổng, đồng thời thông báo cho người dùng về hạn chế của mô hình và đảm bảo tuân thủ các quy định xã hội [50]. Đặc biệt khi chính phủ trên toàn thế giới áp đặt các quy định nghiêm ngặt hơn đối với các hệ thống trí tuệ nhân tạo như Quy định Bảo vệ Dữ liệu Chung của EU (GDPR) và

Khung pháp lý cho Đạo luật Trí tuệ Nhân tạo của EU, nhấn mạnh trách nhiệm giải trình và tính minh bạch trong các hệ thống NLP.

2.2.3 Hiểu biết khoa học về việc phát triển NLP

Việc phát triển các mô hình NLP phức tạp đòi hỏi sự hiểu biết sâu sắc về cả hiện tượng ngôn ngữ và những phức tạp trong học máy. Những hiểu biết từ X-NLP có thể hình thành lý thuyết và nghiên cứu trong ngôn ngữ học và khoa học nhận thức, dẫn đến những khám phá và tiến bộ mới [21]. Hơn nữa, khả năng diễn giải trong việc phát triển NLP giúp xác định lỗi, sai lệch và các khu vực cần cải thiện, đồng thời hướng dẫn tinh chỉnh mô hình để đạt hiệu suất tốt hơn [10]. Điều này không chỉ nâng cao độ bền vững của mô hình mà còn thúc đẩy sự đổi mới, làm sáng tỏ các khía cạnh khác của ngôn ngữ chưa rõ. Ngoài ra, các hệ thống NLP còn đóng vai trò như một công cụ giáo dục trực quan thông qua các kỹ thuật trực quan hóa quá trình xử lý và ra quyết định của mô hình.

2.3 Phạm vi giải thích

Phạm vi giải thích trong X-NLP được chia thành hai khía cạnh chính*: mức độ giải thích* và *nguồn gốc giải thích*.

Mức độ giải thích:

Giải thích cục bộ - cung cấp giải thích cho các dự đoán riêng lẻ dựa trên dữ liệu đầu vào cụ thể.

Giải thích toàn cục - cung cấp cái nhìn tổng quan về quy trình dự đoán của mô hình mà không phụ thuộc vào đầu vào cụ thể.

Giải thích theo lớp - giải thích cách mô hình đưa ra dự đoán cho một lớp cụ thể.

Nguồn gốc giải thích

Giải thích nội tại: được tạo ra đồng thời với dự đoán của mô hình tác vụ bằng việc sử dụng thông tin trung gian trong quá trình dự đoán. Thường gặp ở các mô hình NLP

truyền thống như cây quyết định (decision trees) và các mô hình dựa luật, cũng như các mô hình NLP hiện đại sử dụng cơ chế chú ý. Do đó, mô hình tác vụ cần giải thích và mô hình giải thích nội tại thường là cùng một mô hình.

Giải thích hậu nghiệm: được tạo ra sau khi mô hình đã đưa ra dự đoán thông qua các bước bổ sung, giúp các mô hình có thể giải thích được mà không ảnh hưởng đến kiến trúc hay hiệu suất của chúng. Tuy nhiên, cần đảm bảo mô hình giải thích hoàn toàn tách biệt với mô hình tác vụ, thể hiện chính xác quá trình lập luận của mô hình tác vụ [62].

Kết hợp hai khía cạnh trên sẽ tạo ra bốn phân loại giải thích trong XAI, chi tiết trong ***Bảng 1***.

Bảng 1. *Các phân loại giải thích.*

	Giải thích cục bộ	**Giải thích toàn cục**
Giải thích nội tại	Giải thích cho từng dự đoán được đưa ra trong quá trình ra quyết định.	Toàn bộ nguyên tắc hoặc quá trình suy luận của mô hình tác vụ được ghi lại để giải thích cho bất kỳ dự đoán nào.
Giải thích hậu nghiệm	Sau khi mô hình tác vụ đưa ra dự đoán, thực hiện các bước xử lý bổ sung để giải thích cho từng dự đoán.	Toàn bộ cơ sở lý luận hoặc quá trình suy luận của mô hình tác vụ được tái hiện thông qua thực hiện các thao tác bổ sung nhằm giải thích cho bất kỳ dự đoán nào.

2.4 Chất lượng giải thích

Các giải pháp X-NLP nhằm hướng tới hai tiêu chí chính: Tính dễ hiểu (human-grounded) và tính trung thực (function-grounded) [21]:

Tính dễ hiểu (Human-Grounded): Đưa ra các giải thích đảm bảo dễ hiểu, dễ tiếp cận với người dùng, kể cả những người không có chuyên môn. Sự tham gia của con người là rất quan trọng để đánh giá tính dễ hiểu, cung cấp những hiểu biết có ý nghĩa về cách mô hình ra quyết định. Điều này giúp người dùng đánh giá độ tin cậy của mô hình và hỗ trợ các nhà phát triển gỡ lỗi các dự đoán sai.

Tính trung thực (Function-Grounded): Đảm bảo các giải thích phản ánh trung thực quá trình dự đoán của mô hình. Các giải thích có giá trị cần phơi bày những sai sót trong quá trình ra quyết định của mô hình. Ngay cả các giải thích nội tại như điểm attention cũng không phải lúc nào cũng phản ánh đúng lý luận thực sự của mô hình [33,66].

Việc cân bằng giữa tính dễ hiểu và tính trung thực có thể gặp nhiều thách thức. Các giải thích hướng tới tính trung thực thường phản ánh quá trình ra quyết định phức tạp của mô hình nhưng lại không dễ hiểu đối với người dùng.

Hơn nữa, tính độc lập với mô hình (model-agnosticism) trong các giải thích hậu nghiệm cũng rất cần thiết. Các kỹ thuật giải thích được thiết kế hoàn toàn độc lập với mô hình, không phụ thuộc vào kiến trúc và tham số cụ thể của mô hình cần giải thích. Các giải thích được tạo ra chỉ dựa trên dữ liệu khiến chúng trở nên rất linh hoạt và có thể áp dụng cho nhiều mô hình tác vụ khác nhau.

2.5 Phân loại giải thích

Dựa vào nội dung giải thích, các phương pháp X-NLP được phân loại thành bốn nhóm chính:

Giải thích dựa trên tầm quan trọng đặc trưng: Làm rõ tầm quan trọng của các đặc trưng được sử dụng bởi mô hình tác vụ để hình thành dự đoán (**Chương 3**).

Giải thích dựa trên ví dụ: Sử dụng các ví dụ cụ thể để làm sáng tỏ lý do đằng sau các dự đoán, chẳng hạn như chỉ ra các ví dụ từ dữ liệu huấn luyện mà mô hình tác vụ đã học được.

Giải thích dựa trên chứng minh: Cung cấp kiến thức về quá trình lập luận của mô hình tác vụ, trình bày chuỗi logic liên quan đến các dự đoán của chúng.

Giải thích dựa trên quy nạp khai báo: Cung cấp các phép biến đổi, chẳng hạn như các quy tắc hoặc mã giả (pseudocode), mà các mô hình tác vụ dựa vào khi đưa ra dự đoán.

Đối tượng giải thích sẽ luôn rơi vào một trong hai tình huống sau:

Giải thích mô hình tác vụ: Đối tượng cần giải thích chính là mô hình tác vụ, cùng với các hành vi và dự đoán của mô hình.

Giải thích mô hình thay thế (stuntman model): Một mô hình "stuntman" được thiết kế để có tính giải thích cao hơn so với mô hình tác vụ chính. Mô hình thay thế này phải phản ánh được hành vi của mô hình tác vụ trong phạm vi dữ liệu đang xem xét.

Chương 3: KHẢO SÁT HIỆN TRẠNG

Độ quan trọng của đặc trưng $\mathcal{FI}$ với đầu vào x, dự đoán $\hat{y}$ của mô hình tác vụ $\mathcal{M}$ được đo bằng độ quan trọng của n đặc trưng trong không gian đặc trưng $\mathcal{F}$:

$\mathcal{FI}(x, \hat{y}, \mathcal{M}): \mathcal{F} \to \mathbb{I}^n$ (với $\mathbb{I} \subset \mathbb{R}, \mathbb{I} \in [0,1]$ hoặc $\mathbb{I} \in [-1,1]$ (1)

Các giải thích này chủ yếu áp dụng cho phương pháp giải thích cục bộ, bao gồm cả Hậu nghiệm cục bộ và Nội tại cục bộ. Hướng tiếp cận này phổ biến vì dễ hiểu và phù hợp với cơ chế của các mô hình học máy. Tư liệu để giải thích có thể dễ dàng xác định chẳng hạn như đặc trưng từ, cụm từ, khái niệm, kiến thức ngôn ngữ và các đặc trưng cụ thể cho tác vụ (**Mục 163.3** đến **3.7**).

3.1 Các thước đo tầm quan trọng

Để đánh giá tầm quan trọng của đặc trưng $\mathfrak{f}$, các nghiên cứu thường loại bỏ $\mathfrak{f}$ ra khỏi đầu vào và quan sát sự thay đổi trong đầu ra của mô hình:

$$\mathcal{FI}_{\mathfrak{f}}(x, \hat{y}, \mathcal{M}) = \nabla_{\mathfrak{f} \leftarrow null}\, p(\hat{y}|x; \theta_{\mathcal{M}}) \quad (2)$$

Để không bị ảnh hưởng bởi sự tương tác giữa các đặc trưng, một số nghiên cứu [36,40] đề xuất đánh giá riêng lẻ từng đặc trưng, trong khi các công trình khác tập trung khai thác tầm quan trọng của việc tích hợp nhiều đặc trưng [44,59].

Tính cần (necessity) và *tính đủ* (sufficiency) là hai tiêu chí có thể được sử dụng để đánh giá tầm quan trọng của đặc trưng [6]. *Tính cần* cao cho thấy nếu đặc trưng x thay đổi, dự đoán sẽ thay đổi ngay cả khi các đặc trưng khác được giữ nguyên. Mặt khác, *tính đủ* cao có nghĩa là đầu ra sẽ không thay đổi bất kể có sự thay đổi của các đặc trưng đầu vào khác, miễn là đặc trưng x không thay đổi.

Xét đặc trưng x_i trong mẫu đầu vào x. Giá trị hiện tại của x_i là a và $f(\mathrm{x}) = \mathrm{y}$ là lớp dự đoán đầu ra của x. $\mathcal{NB}_{\mathrm{y(x)}}$ là tập hợp các mẫu đầu vào lân cận x' mà $f(x') = f(x) = y$

và x'_i = a = x_i. Ngược lại, $\mathcal{NB}_{\tilde{y}(x)}$ đại diện cho các mẫu x' mà $f(x') = y' \neq f(x)$ và $x'_i = a' \neq x_i$.

Tính cần của x_i trong việc dự đoán nhãn y là xác suất $f(x') \neq y$ khi x'_i trong $\mathcal{NB}_{y(x)}$ được thay đổi sao cho $x'_i \neq a$, trong khi giữ nguyên các đặc trưng khác trong x':

$$\mathcal{N}_x(x_i, y) = P_{x' \in \mathcal{NB}_y(x)}(f(x') \neq y \mid x'_i \leftarrow a') \qquad (3)$$

Tính đủ của x_i trong việc dự đoán nhãn y, là xác suất f(x') = f(x) = y khi thay đổi từ a' sang a của x'_i trong $\mathcal{NB}_{\tilde{y}(x)}$, trong khi giữ nguyên các đặc trưng khác trong x':

$$\mathcal{S}_x(x_i, y) = P_{x' \in \mathcal{NB}_{\tilde{y}}(x)}(f(x') = y \mid x'_i \leftarrow a) \qquad (4)$$

Các phép đo độ tương đồng hoặc các kỹ thuật biến đổi dữ liệu (**Mục 3.2**) kết hợp với các mô hình ngôn ngữ được sử dụng trong việc chọn ra các mẫu đầu vào x' trong $\mathcal{NB}_{y(x)}$ hoặc $\mathcal{NB}_{\tilde{y}(x)}$ từ bộ ngữ liệu.

3.2 Nhiễu loạn dữ liệu

3.2.1 Các phương pháp

Để ước lượng được tầm quan trọng của vùng dữ liệu bị nhiễu loạn, người ta sử dụng chiến lược thay đổi có chủ đích dữ liệu đầu vào và theo dõi sự thay đổi trong dự đoán của mô hình [15,44,59,65]. LIME (Local Interpretable Model-agnostic Explanations) là một ví dụ điển hình, trong đó từng đặc trưng được loại bỏ ngẫu nhiên từ tập dữ liệu đầu vào $\tilde{X}$, và các dự đoán tương ứng được ghi lại [59]. Dựa trên độ tương đồng $\mathcal{D}$ (tương đồng cosine) với mẫu dữ liệu gốc x, có n đặc trưng, trọng số $w_x(\tilde{x})$ cho mỗi mẫu dữ liệu bị xáo trộn $\tilde{x} \in \tilde{X}$ được xác định:

$$w_x(\tilde{x}) = \sqrt{\exp\left(\frac{-\mathcal{D}(x,\tilde{x})}{(0.75*\sqrt{n})^2}\right)} \qquad (5)$$

Việc gán trọng số này biến $\widetilde{\mathcal{X}}$ thành phân phối tuyến tính. Sau đó, LIME huấn luyện mô hình hồi quy tuyến tính sử dụng tập dữ liệu có trọng số này, và các hệ số β_i trong phương trình hồi quy (6) đại diện cho tầm quan trọng của mỗi đặc trưng f_i bị loại bỏ trong $\tilde{x}_i$:

$$g(\tilde{x}) = \beta_0 + \beta_1\tilde{x}_1 + \beta_2\tilde{x}_2 + \cdots + \beta_n\tilde{x}_n \qquad (6)$$

Như một mô hình thay thế, LIME sử dụng hàm mất mát bình phương có trọng số trong phương trình (7) để giải thích các mô hình phức tạp bằng một mô hình hồi quy tuyến tính đơn giản:

$$\mathcal{L}(\mathcal{M}, g, w_x) = \sum_{\tilde{x},\widetilde{x'}\in\tilde{X}} w_x(\tilde{x})\,(\mathcal{M}(\tilde{x}) - g(\widetilde{x'}))^2 \qquad (7)$$

Biến thể của LIME, LIMSSE, lấy các chuỗi con có độ dài cố định từ dữ liệu đầu vào làm đặc trưng nhiễu loạn [56]. Điều này giúp đảm bảo thứ tự từ trong câu và cải thiện hiệu quả trong lĩnh vực NLP.

SHAP [44] sử dụng Lý thuyết trò chơi, một phương pháp tiên tiến để xác định giá trị Shapley cho mỗi đặc trưng. So với LIME, nó cung cấp các giải thích toàn diện và phù hợp với lập luận của con người, nhờ quan tâm đến tương tác giữa các đặc trưng. Tuy nhiên, việc tính toán Shapley với số lượng đặc trưng lớn có thể tốn kém. SHAP đề xuất kết hợp Shapley với phương trình tuyến tính của LIME để xấp xỉ giá trị Shapley:

$$w_x(\tilde{x}, \overrightarrow{cv}) = \frac{(n-1)}{\binom{n}{|\overrightarrow{cv}|}|\overrightarrow{cv}|(n-|\overrightarrow{cv}|)} \qquad (8)$$

Trong đó $\overrightarrow{cv}$ là vector one-hot biểu diễn các đặc trưng bị nhiễu loạn trong $\tilde{x}$, và $|\overrightarrow{cv}|$ là số lượng giá trị 1 trong $\overrightarrow{cv}$. Trọng số này giúp xấp xỉ giá trị Shapley ϕ_i của đặc trưng thứ *i* khi huấn luyện mô hình hồi quy tuyến tính:

$$g(\tilde{x}) = \phi_0 + \sum_{j=1}^{n} \phi_j\,\widetilde{x_j} \qquad (9)$$

Nhiễu loạn dữ liệu cũng được sử dụng để đánh giá khả năng giải thích của mô hình bằng cách che giấu tuần tự các token nổi bật (độ quan trọng cao) trong đầu vào và lấy trung bình các thay đổi trong đầu ra mô hình [84]. Mô hình tác vụ nào có điểm số trung bình của sự thay đổi cao hơn được xem là dễ hiểu hơn [83].

3.2.2 Thảo luận

Mặc dù LIME và SHAP cung cấp lời giải thích gần với lập luận con người, nhưng gây ra lo ngại về tính chính xác của mô hình hồi quy tuyến tính trong việc bắt chước hành vi của mô hình hộp đen. Hơn nữa, LIME không được thiết kế riêng cho NLP, vì vậy mô hình hồi quy tuyến tính của nó không hoàn toàn phù hợp với dữ liệu văn bản, đặc biệt là khi so sánh với các mô hình học sâu như LSTM, CNN và RNN.

Ngoài ra, không thể đảm bảo rằng dữ liệu nhiễu loạn và dữ liệu gốc truyền đạt cùng một ý nghĩa. Để khắc phục điều này, một đề xuất sử dụng các biểu diễn ẩn để tạo ra dữ liệu nhiễu loạn giúp duy trì sự tương đồng về ngữ nghĩa [3]. Hơn nữa, việc xóa hoặc che giấu các token cũng gây ra những đánh giá không khách quan về tầm quan trọng của đặc trưng, đặc biệt khi áp dụng trong các mô hình ngôn ngữ được huấn luyện để điền vào chỗ trống hoặc dự đoán từ tiếp theo.

Tác động của nhiễu loạn đầu vào lên dự đoán của mô hình là đặc thù cho từng mô hình, đặt ra lo ngại về độ tin cậy của kỹ thuật nhiễu loạn dữ liệu để so sánh tính giải thích giữa các mô hình khác nhau. LISA (Tích lũy Ngữ nghĩa Tầng lớp) [30] ra đời như giải pháp thay thế để đánh giá tác động của mỗi token lên dự đoán. Tuy nhiên, LISA chỉ nắm bắt các token quan trọng nhưng không liên tục trong câu.

3.3 Đặc trưng ngôn ngữ

3.3.1 Các phương pháp

Phương pháp giải thích dựa vào đặc điểm ngôn ngữ được trích xuất từ việc thăm dò cấu trúc – structural probing (tức phân tích vector biểu diễn trong các mô hình học sâu).

Điều này khác với thăm dò hành vi – behavioral probing (đánh giá phản hồi đối với đầu vào cụ thể - **Mục 3.2**), sử dụng các mạng nơ-ron đơn giản như mạng nơ-ron nhiều lớp hoặc mô hình hồi quy logistic để khai thác thông tin ngôn ngữ từ trong các lớp biểu diễn trung gian của mô hình, điều chỉnh các vector biểu diễn từ và câu để phù hợp với các cấu trúc ngôn ngữ. Biểu diễn mức từ giúp khai thác các đặc trưng loại từ, thể và giọng của từ [16,65], trong khi biểu diễn câu cung cấp thông tin về độ dài câu, thứ tự từ và sự hiện diện của từ [1].

Nghiên cứu BERTology áp dụng công cụ thăm dò này vào BERT để đánh giá tầm quan trọng của các lớp đối với độ chính xác của dự đoán cho các khía cạnh cú pháp khác nhau, như mối quan hệ thành phần và phụ thuộc [70]. Các đầu chú ý (attention heads) trong cùng một lớp thường thể hiện các kiểu hành vi tương tự, như tập trung vào cùng vị trí hoặc phân bổ sự chú ý đồng đều trên các từ, theo nghiên cứu [14,60]. Ngoài ra, [14] còn cho rằng một số nhóm đầu chú ý chuyên biệt hóa trong việc xác định các thành ngữ cú pháp cụ thể với độ chính xác cao (trên 75%), như động từ, danh từ, giới từ và đại từ sở hữu. Tuy nhiên, do sự phức tạp của toàn bộ cây cú pháp, không một đầu chú ý đơn lẻ có thể nắm bắt toàn bộ thông tin của nó.

Loại từ (Part Of Speech - POS) là thuộc tính quan trọng trong việc phân tích nguồn gốc của các dự đoán. Ví dụ, trong nghiên cứu [39] liên quan đến bệnh Alzheimer, các bản ghi phỏng vấn được đánh nhãn bằng các thẻ POS và đưa vào mạng học sâu n nơ-ron. Để biểu diễn các bản ghi này, họ sử dụng giá trị kích hoạt của các nơ-ron như các tọa độ trong không gian n chiều. Kỹ thuật phân cụm K-means được áp dụng vào các vector này cho thấy các cụm tiết lộ triệu chứng của bệnh Alzheimer có phân bố thẻ POS tương tự, trong khi các cụm không có triệu chứng Alzheimer lại có các mẫu thẻ POS khác biệt. Họ đã phát triển một chương trình để tự động phân tích các mẫu POS này như một giải thích cho việc phân tích một bản ghi là có chỉ ra dấu hiệu của bệnh Alzheimer hay không.

3.3.2 Thảo luận

Cách diễn giải với người dùng về các giải thích dựa trên đặc trưng ngôn ngữ sẽ quyết định liệu chúng có dễ hiểu hay không. Tuy nhiên, phương pháp này chủ yếu phù hợp với các chuyên gia, những người có thể đánh giá hiệu quả các lời giải thích này.

Thăm dò cấu trúc cho phép trực tiếp đào sâu vào các biểu diễn ẩn của mô hình tác vụ, phù hợp với tính trung thực của mô hình [16]. Đây là kỹ thuật hậu nghiệm, áp dụng trên các mô hình tiền huấn luyện khác nhau mà không cần điều chỉnh chúng. Hơn nữa, những kiến thức trong các biểu diễn của mô hình có thể được chuyển giao cho các mô hình tương tự, tăng tính linh hoạt của kỹ thuật này.

Phương pháp *BERTology* cung cấp một kỹ năng diễn giải tinh tế và chuyên biệt, hữu ích cho việc tinh chỉnh mô hình và hỗ trợ học chuyển giao [70]. Tuy nhiên, nó đòi hỏi nhiều tài nguyên tính toán và việc sử dụng cơ chế attention cho giải thích gây tranh cãi trong cộng đồng khoa học (**Mục 3.8**). Do đó, dù phương pháp này khám phá được các vùng tập trung cú pháp của BERT, nhưng không phải lúc nào cũng cung cấp cái nhìn sâu sắc về cách sự tập trung này ảnh hưởng đến dự đoán của mô hình.

Hơn nữa, cần cẩn trọng với các vector biểu diễn có số chiều lớn, vì các mô hình thăm dò có thể suy luận nhiều hơn từ các vector này so với những gì mà mô hình tác vụ thực sự sử dụng cho quyết định, đe dọa tính trung thực của mô hình. Hơn nữa, thăm dò hiệu quả đòi hỏi các tài nguyên ngôn ngữ như tập dữ liệu chú thích và từ điển, phải nắm bắt chính xác các đặc điểm ngôn ngữ cần quan tâm [16,71]. Do đó, chất lượng của lời giải thích phụ thuộc vào sự sẵn có và chất lượng của các tài nguyên này, đòi hỏi một lượng lớn thời gian và công sức để có thể thu thập được.

Phân cụm K-means [39] là công cụ giải thích hữu ích nhưng không hoàn toàn bao quát được sự phức tạp của ngôn ngữ và sự liên quan của nó đến các tác vụ NLP hiện tại. Do đó, nên xem phân cụm là thành phần của một khung giải thích rộng hơn thay vì là một giải pháp toàn diện. Ngoài ra, độ nhạy cảm của thuật toán đối với các giá trị ngoại

lai (outliers) có thể làm méo mó kết quả, đặt ra câu hỏi về độ chính xác trong việc phản ánh logic nội tại của mô hình tác vụ (tính trung thực).

3.4 Đặc trưng khái niệm

3.4.1 Các phương pháp

Trong nghiên cứu X-NLP, các phương pháp liên quan đến các khía cạnh khái niệm chủ yếu nhắm đến hai mục tiêu.

Thứ nhất, giảm thiểu thiên vị trong các mô hình học máy. Sự thiên vị thường xảy ra giữa các tính từ miêu tả các nhân vật chính trị [25] hoặc giữa vai trò nghề nghiệp dựa trên giới tính [34], chẳng hạn mô hình sẽ hiểu bác sĩ luôn là giới tính nam.

Thứ hai, xác định tác động của các khái niệm quan trọng. Các nghiên cứu này khám phá cách các khái niệm như màu da, sắc tộc và tôn giáo ảnh hưởng đến thuật toán phát hiện ngôn từ căm thù, đặc biệt trong bối cảnh gia tăng thái độ kì thị người Châu Á trong đại dịch COVID-19 [54].

Để đo lường độ nhạy cảm của mô hình đối với các khái niệm đầu vào, nghiên cứu [54] sử dụng phương pháp Kiểm tra Vector Kích hoạt Khái niệm (TCAV), từ lĩnh vực thị giác máy tính. TCAV xây dựng các vector kích hoạt cho một khái niệm bằng cách lấy trung bình các kích hoạt của các mẫu ngẫu nhiên của khái niệm và lặp lại với các tập mẫu khác nhau để tạo ra nhiều vector cho mỗi khái niệm. Độ nhạy cảm của mô hình đối với một khái niệm được định lượng bằng sự thay đổi trong giá trị xác suất (logit) khi một phần nhỏ của vector kích hoạt được nối vào vector của mẫu đầu vào.

Cơ chế attention phổ biến trong các mô hình tác vụ, tính toán điểm số attention để xác định mức độ liên quan giữa các khái niệm với dự đoán. [81] đã minh họa điều này trong tác vụ hoàn thiện cơ sở tri thức (tức tìm kiếm các mối quan hệ mới giữa các cặp khái niệm trong văn bản, sau đó thêm chúng vào cơ sở tri thức), nơi mà điểm số attention tiết lộ các khái niệm quan trọng ảnh hưởng đến các mối quan hệ được phát hiện. [34] sử

dụng các phân tích hòa giải nhân quả [37] để theo dõi sự lan truyền thiên lệch giới tính trong mô hình dự đoán như GPT-2, xác định các khái niệm và đầu attention cụ thể làm tăng cường thiên lệch.

Bên cạnh các khái niệm văn bản, cơ chế attention cũng áp dụng cho các khái niệm trong cơ sở tri thức. [68] phát triển mô hình giải thích cho tác vụ hỏi đáp sử dụng cơ chế chú ý đa bước (multi-hop), tính toán điểm số attention cho các khái niệm trong cơ sở tri thức. Một khái niệm trong cơ sở tri thức bao gồm hai thực thể (*s* và *o*) và một mối quan hệ (*r*). Điểm số attention cho các khái niệm này được tính bằng cách khởi tạo vector ngữ cảnh c_0 từ truy vấn q và cập nhật c_t ở các bước t:

$$c_t = W_t(c_{t-1} + W_p \sum_{(k,v)\in\mathcal{M}} softmax(c_{t-1} \cdot k)\, v) \quad (10)$$

Trong phương trình trên, W_t và W_p là ma trận trọng số được huấn luyện. Kết quả ở bước t cuối cùng được đưa vào lớp fully-connected để tính ra vector $b \in \mathbb{R}_d$. Sau đó, tích vô hướng giữa vector b và các nhúng thực thể cho ra điểm số attention cho truy vấn q liên quan đến mỗi thực thể.

Tuy nhiên, một hạn chế đáng chú ý của cơ chế attention là mỗi cặp khái niệm chỉ có duy nhất một mối quan hệ, vấn đề này được giải quyết bằng cách chia nhỏ ma trận attention thành các ma trận con tương ứng với các mối quan hệ khác nhau giữa hai khái niệm, nâng cao khả năng giải thích của mô hình [38].

Phương pháp Xáo trộn dữ liệu như CausaLM [25] sử dụng phân tích cảm xúc bằng việc che giấu các khái niệm để đánh giá tác động đến dự đoán của mô hình. Không giống LIME, che giấu dữ liệu một cách trực tiếp, CausaLM thêm các thuật ngữ phủ định vào hàm mất mát của mô hình ngôn ngữ để "lãng quên" các khái niệm dự kiến được che giấu trong quá trình huấn luyện, đồng thời bảo toàn các khái niệm khác bằng một thuật ngữ kiểm soát tích cực:

$$\mathcal{L}(\theta_{LM}, \theta_t, \theta_{cc}, \theta_{tc}) = \frac{1}{n}\left(\sum_{i=1}^{n} \mathcal{L}_t^i(\theta_{LM}, \theta_t) + \sum_{i=1}^{n} \mathcal{L}_{cc}^i(\theta_{LM}, \theta_{cc}) - \lambda \sum_{i=1}^{n} \mathcal{L}_t^i(\theta_{LM}, \theta_{tc})\right) \quad (11)$$

Ở đây, θ_{LM} là tham số của toàn bộ mô hình ngôn ngữ, ngoại trừ θ_t, θ_{cc}, và θ_{tc} lần lượt là các tham số của tác vụ chính (downstream task), các *khái niệm kiểm soát* (*cc*) cần được giữ lại và các *khái niệm xử lý (tc*) được che giấu. Phương pháp này giúp CausaLM vừa là mô hình tác vụ vừa là mô hình giải thích, nghiêng về giải thích nội tại hơn là các giải thích hậu nghiệm.

3.4.2 Thảo luận

Cơ chế attention trong giải thích AI dựa trên khái niệm gây tranh cãi về tính hợp lệ [34] (**Mục 3.8**). Nghiên cứu [68] chỉ ra rằng các giải thích hậu nghiệm vượt trội hơn hẳn các giải thích nội tại dựa trên attention, nhấn mạnh tính hiệu quả của attention như một phương pháp giải thích độc lập.

TCAV [54] có tính toàn diện hơn nhờ vào việc lấy mẫu rộng rãi các trường hợp khái niệm trên nhiều tập dữ liệu. Bằng việc lấy trung bình của nhiều vector kích hoạt, TCAV cung cấp một thước đo đáng tin cậy về độ nhạy cảm của đầu ra đối với các khái niệm cụ thể. Tuy nhiên, việc vay mượn trực tiếp các kỹ thuật thị giác máy tính vào NLP có thể không phù hợp do sự khác biệt về cấu trúc dữ liệu. Giả định rằng nhúng khái niệm trong câu chỉ bằng cách thêm vector kích hoạt trung bình là đơn giản hóa quá mức và cần nghiên cứu sâu hơn.

CausaLM [25] giới thiệu một cách tiếp cận mới bằng cách xáo trộn dữ liệu để đánh giá đặc trưng khách quan hơn. Tuy nhiên, kỹ thuật can thiệp vào hàm mất mát của mô hình, đặc thù cho mô hình và tác vụ, làm giới hạn tính ứng dụng và ảnh hưởng đến tính trung thực của mô hình tác vụ. Do đó, cần có những nghiên cứu sâu hơn để khắc phục các hạn chế này.

3.5 Đặc trưng cụm từ

3.5.1 Đặc trưng cụm từ cấu trúc phẳng

Khi đánh giá cụm từ như một cấu trúc phẳng, một kỹ thuật đề xuất sử dụng cửa sổ trượt với độ rộng *w* di chuyển dọc theo câu để trích xuất các chuỗi con có độ dài *w* làm đặc trưng [4]. Tương tự LIMSSE (**Mục 3.2.1**) trong việc chọn đặc trưng, nhưng khác biệt ở chỗ cung cấp giải thích nội tại trực tiếp liên quan tới hành vi của mô hình, không phải là phép xấp xỉ. Các chuỗi con tốt nhất được chọn qua lớp max-over-pooling. Sau đó, lớp fully-connected softmax sẽ tính toán xác suất nhãn và cung cấp dự đoán kèm theo giải thích về các chuỗi con quan trọng.

Trong tác vụ phân loại hồ sơ y tế, CAML (Convolutional Attention for Multi-Label classification) [52] sử dụng cơ chế attention trên cụm 4 từ (4-gram) để tính toán vector attention cho mỗi mã ICD (International Classification of Diseases), xác định các cụm từ quan trọng trong hồ sơ mà mô hình phân loại dựa vào.

Đối với tầm quan trọng của các khung dựa trên cơ chế attention, [12] xây dựng một mô hình tác vụ với hai bộ dự đoán: bộ dự đoán lý luận và bộ dự đoán phản lý. Họ tối ưu mô hình bằng cách tối đa hóa chỉ số F1 cho bộ dự đoán lý luận và tối thiểu hóa cho bộ dự đoán phản lý. Việc tối ưu kép theo hai hướng đối lập nhau giúp mô hình đạt được cơ chế chú ý gần với con người.

3.5.2 Đặc trưng cụm từ cấu trúc phân cấp

Khi xem cụm từ như cấu trúc phân cấp, giải thích phân cấp cho cụm từ phải cung cấp thông tin về tầm quan trọng không phụ thuộc vào ngữ cảnh của mỗi cụm con hoặc token và sự kết hợp phần dư. Các mô hình như LIME và SHAP mắc sai lầm khi đánh giá các từ trong ngữ cảnh cụ thể mà không phản ánh đúng sự kết hợp ngữ nghĩa, không phản ánh được sự hiểu biết như con người. Ngược lại, giải thích phân cấp xác định rõ cảm xúc của từ trong ngữ cảnh cụ thể, giúp định lượng hiệu ứng kết hợp.

Các kỹ thuật truyền thống như LIME hoặc SHAP gặp phải hạn chế khi không cô lập ngữ cảnh trong việc đánh giá tầm quan trọng của đặc trưng. Để cung cấp giải thích phân cấp, cần ước lượng tầm quan trọng độc lập với ngữ cảnh trước khi đánh giá sự kết hợp phần dư. Tầm quan trọng độc lập với ngữ cảnh của một cụm từ hoặc token có thể được tính toán thông qua các thuật toán như Contextual Decomposition (CD), Agglomerative Contextual Decomposition (ACD) [64], Sampling and Contextual Decomposition (SCD) [36], Sampling and Occlusion (SOC) [36,40].

Dù có nhiều biến thể nhưng các thuật toán đều có chung một nguyên tắc cơ bản là phân tách trạng thái ẩn *h* của mô hình tác vụ thành hai thành phần: β, biểu diễn tầm quan trọng độc lập với ngữ cảnh; và γ, đại diện cho ảnh hưởng ngữ cảnh còn lại, thỏa phương trình (12):

$$h = \beta + \gamma \quad (12)$$

Quá trình phân tách này được áp dụng đồng nhất từ lớp đầu vào đến lớp đầu ra. Đối với hàm kích hoạt tuyến tính $h = W_i x_t + b_i$, với β là $W_i x_t$. Đối với hàm kích hoạt phi tuyến tính $h' = \sigma(h)$, CD tính toán đóng góp duy nhất từ cụm t bằng cách đo lường sự khác biệt trung bình trong hàm kích hoạt khi giả định rằng có hay không có sự tác động của ngữ cảnh (ζ là độ lệch):

$$\beta = \frac{1}{2}[\sigma(\beta + \gamma + \zeta) - \sigma(\gamma + \zeta)] + \frac{1}{2}[\sigma(\beta + \zeta) - \sigma(\zeta)] \quad (13)$$

3.5.3 Thảo luận

Các phương pháp trên nhấn mạnh việc phân tích các đoạn và cụm từ trong văn bản, cung cấp cái nhìn sâu sắc hơn về ngữ nghĩa và ngữ pháp so với các kỹ thuật chỉ dựa trên từng token riêng lẻ. Giải thích cụm từ phẳng tuy đơn giản dễ hiểu nhưng thiếu chi tiết về sự kết hợp phần dư.

Phương pháp [4] sử dụng max-over-pooling để giảm kích thước của ma trận đặc trưng, nâng cao hiệu quả tính toán. Tuy nhiên, sự phụ thuộc vào phương pháp kỹ thuật

cửa sổ trượt làm tiềm ẩn tính chủ quan và có thể bỏ sót thông tin quan trọng. Quá trình tối ưu kép [12] giúp đạt được cơ chế chú ý gần với con người nhưng phức tạp và tốn kém chi phí tính toán.

Giải thích cụm từ phân cấp, lý tưởng cho các chuyên gia, hiệu quả cho việc sửa lỗi nhưng hạn chế bởi tính đặc thù trong tác vụ. Phương pháp này được phát triển với mục đích phân tích cảm xúc nhưng không phù hợp cho các mô hình NLP cổ điển hoặc kiến trúc khác. Ngoài ra, phương pháp CD [53] và phương pháp ACD [64] không thỏa mãn tính chất độc lập ngữ cảnh của β vì việc tính β phần nào liên quan đến γ, phụ thuộc vào ngữ cảnh. Do đó, các phương pháp SCD và SOC [36,40] cung cấp sự phân tách chính xác hơn, hạn chế tác động của ngữ cảnh.

3.6 Đặc trưng token

Token là đặc trưng bề mặt rõ ràng nhất trong NLP, dễ hiểu và dễ tiếp cận với người dùng. Giải thích các dự đoán của mô hình NLP thường tập trung vào tầm quan trọng của token, thu hút nhiều sự chú ý từ cộng đồng nghiên cứu.

3.6.1 Huấn luyện nhận dạng các token quan trọng

Sử dụng tập dữ liệu giải thích có chú thích cùng trọng số độ quan trọng của đặc trưng giúp nâng cao độ tin cậy của các giải thích từ mô hình NLP. Công trình [61] đã cải thiện hiệu suất của mô hình bằng cách tận dụng các giải thích "vàng" từ tập dữ liệu ấy để tối ưu hóa hàm mất mát, mã hóa các giải thích thành một vector one-hot, kí hiệu là A, với giá trị 1 tương ứng đặc trưng không liên quan đến dự đoán. Vector A được tích hợp vào hàm mất mát của mô hình NLP, nhằm giảm thiểu độ dốc (gradient) đầu ra cho các đầu vào có giá trị 1 trong vector A, được trình bày trong phương trình (14) đến (17).

$$RightPrediction = \sum_{n=1}^{N} \sum_{k=1}^{K} -y_{nk} log(\hat{y}_{nk}) \quad (14)$$

$$RightExplanation = \lambda_1 \sum_{n=1}^{N} \sum_{d=1}^{D} (A_{nd} \frac{\partial}{\partial x_{nd}} \sum_{k=1}^{K} log(\hat{y}_{nk}) \quad (15)$$

$$RegularTerm\ = \lambda_2 \sum_i \theta_i^2 \text{ (16)}$$

$$L(\theta, X, y, A) = RightPrediction + RightExplanation + RegularTerm \text{ (17)}$$

Trường hợp không tồn tại giải thích "vàng", vector được khởi tạo với tất cả giá trị 0 và cập nhật trong quá trình huấn luyện mô hình. Công trình [31] giải quyết vấn đề không tồn tại giải thích "vàng" bằng cách đề xuất mô hình học đối kháng không giám sát, huấn luyện đồng thời 3 mô hình: bộ chọn chính, bộ chọn đối kháng và mô hình tác vụ. Bộ chọn chính chọn các token cho tập lý luận từ mẫu đầu vào x, trong khi bộ chọn đối kháng sẽ lựa chọn một tập hợp ngẫu nhiên các token, gọi là tập lý luận đối kháng. Mô hình tác vụ chính sử dụng cả hai tập lý luận trên để đưa ra các dự đoán riêng biệt hoàn toàn từ mỗi tập. Hàm mất mát được thiết kế để đưa ra dự đoán từ tập lý luận gần với nhãn "vàng" hơn những dự đoán từ tập lý luận đối kháng.

3.6.2 Độ quan trọng của token dựa trên cơ chế attention

Giải thích dựa trên cơ chế attention xác định tầm quan trọng nội tại của đặc trưng trong giải thích cục bộ. Điểm attention chỉ ra các token được ưu tiên trong dự đoán [2,9,69]. Trong việc thăm dò không gian ẩn, điểm số attention cũng được sử dụng để giải thích các mô hình tác vụ tạo ra các biểu diễn cho các từ nằm ngoài bộ từ vựng, tiết lộ các từ ngữ cảnh mà nó tập trung vào nhiều nhất [26]. Hơn nữa, [82] giới thiệu các sự kiện attention, cho thấy cách mô hình ngôn ngữ sử dụng các sự kiện văn bản để phân loại, mặc dù nó vẫn tập trung vào các token cụ thể đại diện cho các sự kiện.

Công trình [9] đã tận dụng token [CLS] trong BERT để tính toán độ quan trọng của token trong phân loại văn bản, khai thác vector ngữ cảnh attention của token [CLS] từ các lớp biểu diễn trung gian, tạo ra giải thích về tầm quan trọng của đặc trưng. Ngoài ra, cơ chế attention cũng được sử dụng giữa token trong các văn bản khác nhau, giúp giải thích các phản hồi của chatbot [43] và Suy luận Ngôn ngữ Tự nhiên (NLI). Các mô hình như X-NLP tính toán ma trận attention giữa các từ trong các văn bản khác nhau, nắm bắt các liên kết ngữ nghĩa và cung cấp giải thích với ngữ cảnh rộng hơn.

Bên cạnh đó, cơ chế attention còn được áp dụng trong tác vụ tạo hình ảnh từ văn bản thông qua mô hình khuếch tán ẩn (latent diffusion model). Cơ chế này thiết lập liên kết giữa các bộ nhúng văn bản của token x và các biểu diễn ẩn tọa độ của hình ảnh, gán điểm attention cho mỗi cặp token-patch ảnh để chỉ ra từ nào trong văn bản đầu vào ảnh hưởng đến vùng cụ thể nào trong hình ảnh. Ví dụ được minh họa ở ***Hình 3***.

Hình 3. *Một hình ảnh được tổng hợp và các điểm ảnh nổi bật tương ứng cho ba token "monkey", "hat" và "walking" từ văn bản gợi ý "monkey with hat walking."*

3.6.3 Thảo luận

Các giải thích từ phương pháp phân biệt tập lý luận đối kháng [31] và các giải thích "vàng" không chỉ được tận dụng để huấn luyện mô hình mà còn giúp đưa ra dự đoán, đảm bảo tính trung thực và gần với tư duy của con người do được chú thích dựa trên chuyên môn của con người. Tuy nhiên, hướng tiếp cận này hạn chế trong việc đánh giá tầm quan trọng của đặc trưng bởi chúng chỉ được phân loại theo kiểu nhị phân là quan trọng hoặc không quan trọng mà thiếu đi chiều sâu về sự hiểu biết của con người. Bên cạnh đó, trong các giải thích hậu nghiệm cần cẩn trọng khi sử dụng tập huấn luyện được chú thích bởi con người, vì có thể không phản ánh được quá trình lập luận của mô hình tác vụ làm ảnh hưởng đến tính trung thực.

Vốn dĩ cơ chế attention là một phần trong kiến trúc của mô hình tác vụ, đóng vai trò như "kim chỉ nam" cho quá trình ra quyết định của mô hình nhưng không hoàn toàn phản ánh chính xác độ quan trọng của token [33], được đề cập thêm ở **Mục 3.8**

Đối với các giải thích dựa vào token [CLS], vốn dĩ [CLS] được thiết kế để nắm bắt thông tin đầu vào nên việc giải thích dựa vào vector ngữ cảnh của nó là một biểu diễn

hợp lý nhưng tập trung đơn thuần vào vector biểu diễn của token ở lớp cuối cùng có thể bỏ lỡ các thông tin quan trọng ở các lớp trước và không nắm bắt đầy đủ các phụ thuộc phức tạp.

Ngoài ra, việc kết hợp cơ chế attention giữa token trong các văn bản khác nhau có thể nắm bắt được các liên kết ngữ nghĩa giữa các văn bản, do đó cung cấp được lời giải thích với ngữ cảnh rộng hơn [43,72]. Tuy nhiên, công trình [72], hàm mất mát gây ra cơ chế attention dàn trải đều, khiến mô hình bỏ sót thông tin sắc thái.

Công trình tạo hình ảnh từ văn bản [69] chủ yếu phù hợp với các tác vụ tạo hình ảnh và đòi hỏi nhiều tài nguyên tính toán, giới hạn khả năng sử dụng cho các giải thích nhanh chóng hoặc theo thời gian thực, đồng thời dễ gặp hiện tượng quá khớp (overfitting).

Tóm lại, các dự đoán NLP thường bị ảnh hưởng bởi nhiều hơn các token riêng lẻ trong văn bản. Mặc dù tầm quan trọng của token cung cấp cái nhìn sâu sắc về các từ cụ thể có liên quan nhưng lại không nắm bắt được các cấu trúc ngữ pháp và ngữ nghĩa rộng hơn. Chính vì thế, việc chỉ dựa vào cơ chế attention ở cấp độ token có thể bỏ lỡ những khía cạnh sâu sắc trong hành vi của mô hình.

3.7 Đặc trưng khác

3.7.1 Các phương pháp

Một số nhiệm vụ NLP hưởng lợi từ các đặc trưng đầu vào chuyên biệt liên quan đến từng nhiệm vụ cụ thể. Chẳng hạn, trong việc phát hiện ngôn từ thù địch, các đặc điểm xã hội của tác giả như giới tính, quốc tịch và vùng lãnh thổ là rất quan trọng đối với độ chính xác dự đoán của mô hình [51]. Việc tích hợp một tập hợp đa dạng các đặc điểm xã hội cho phép mô hình phân tích toàn diện hơn về nguyên nhân cơ bản của ngôn từ thù địch, vượt ngoài ngữ cảnh của văn bản. Nghiên cứu đã chỉ ra rằng việc làm giàu mô hình phát hiện ngôn từ thù địch bằng các đặc điểm xã hội đã cải thiện điểm F1 lên 2%. Họ

cũng sử dụng khung SHAP để tính giá trị Shapley, cung cấp giải thích cục bộ định lượng tác động của từng đặc điểm xã hội lên từng dự đoán mô hình.

Việc sử dụng các tài nguyên ngôn ngữ có thể làm phong phú thêm quá trình đưa ra các lời giải thích với các đặc điểm cấu trúc phức tạp. Chẳng hạn, [68] tận dụng cơ sở tri thức làm nguồn tài nguyên ngôn ngữ để làm sáng tỏ các đầu ra của một hệ thống hỏi đáp bằng cách sử dụng các sự kiện trong cơ sở tri thức làm đặc trưng. Mỗi sự kiện mô tả một mối quan hệ r giữa các chủ thể s và đối tượng o, bao gồm key (biểu diễn nhúng của s và r) và $value$ (biểu diễn nhúng của o). Các truy vấn, dưới dạng các câu bị khuyết chủ thể và đối tượng, được xử lý bởi Bi-LSTM và được đưa vào hệ thống hỏi đáp, dự đoán đối tượng còn thiếu từ một tập M các sự kiện trong cơ sở tri thức liên quan đến câu truy vấn. Mô hình giải thích hậu nghiệm áp dụng LIME, che giấu các sự kiện trong M để tạo ra các mẫu nhiễu loạn cho mỗi truy vấn, mỗi sự kiện có 50% khả năng bị che giấu theo phân phối Bernoulli.

Đặc trưng không chỉ bao gồm dữ liệu đầu vào và các tài nguyên ngôn ngữ mà còn bao gồm các thành phần trong kiến trúc mô hình, cùng với các giải thích nhằm đánh giá tác động của chúng đến đầu ra. [23] nghiên cứu sự lan truyền của các gradient đầu vào dọc theo đường đi trên đồ thị tính toán, từ đó cho thấy tầm quan trọng của các thành phần khác nhau trong mạng đối với dự doán dựa trên việc chúng nhận được nhiều hay ít gradient. Để giải quyết tình trạng không khả thi của việc tìm kiếm toàn diện qua các đường con trong đồ thị tính toán, [23] đã sáng tạo ra phương pháp sử dụng lan truyền ngược trong quá trình huấn luyện để thu thập các thống kê gradient của mạng mà không cần đánh giá qua tất cả các đồ thị con. Phương pháp này giới thiệu hai loại *semiring* mới: *max-product semiring*, xác định các đường dẫn gradient tác động mạnh nhất, và *entropy semiring*, định lượng sự phân phối gradient trên toàn mạng. [23] đã tái định nghĩa lan truyền ngược như một bài toán tìm đường đi ngắn nhất, thay thế các phép toán tiêu chuẩn bằng các *semiring* này để tính toán thống kê liên quan đến gradient một cách hiệu quả.

3.7.2 Thảo luận

[68] đưa ra giải thích dựa trên các sự kiện cho phép một cấu trúc phức tạp hơn trong các giải thích, làm chúng trở nên đáng tin cậy hơn vì được hỗ trợ bởi thông tin thực tế trong cơ sở tri thức. Mặc dù phương pháp này tạo ra dữ liệu nhiễu thông qua thao tác che giấu, nhưng nó không chịu ảnh hưởng từ việc giảm chất lượng đánh giá vì nó che giấu các sự kiện trong cơ sở tri thức (thay vì các token) nên mô hình tác vụ không thể suy luận được. Tuy nhiên, phương pháp này phụ thuộc rất lớn vào sự sẵn có và chất lượng của tài nguyên ngôn ngữ, cụ thể là cơ sở tri thức. Nếu cơ sở tri thức không đầy đủ, lỗi thời hoặc không chính xác, có thể dẫn đến các giải thích sai lệch hoặc gây hiểu lầm. Hơn nữa, các đặc điểm sự kiện có thể gây khó hiểu đối với người dùng thông thường. Điều này có thể được khắc phục bằng cách biến đổi các sự kiện thành các câu văn dễ hiểu trong ngôn ngữ tự nhiên, tuy nhiên điều này không đơn giản và đòi hỏi thêm nhiều nỗ lực nghiên cứu.

Việc kết hợp các đặc điểm xã hội vào các mô hình phát hiện ngôn từ thù địch là một bước cải tiến để nâng cao hiệu suất của mô hình, nhưng mang lại rủi ro về quyền riêng tư khi sử dụng dữ liệu cá nhân như vùng lãnh thổ hoặc giới tính. Điều này đòi hỏi các biện pháp bảo vệ quyền riêng tư nghiêm ngặt và các thỏa thuận rõ ràng với người dùng, tránh việc củng cố hoặc tạo ra các định kiến, có thể dẫn đến các kết quả thiên vị hoặc phân biệt đối xử.

Việc nhấn mạnh vai trò của các thành phần mô hình trong dự đoán đầu ra cho phép các nhà nghiên cứu tối ưu hóa và cải thiện các thành phần quan trọng cụ thể, cũng như chẩn đoán các vấn đề tiềm ẩn như xác định các nút thắt, qua đó nâng cao hiệu suất tổng thể của mô hình [23]. Tuy nhiên, sự phụ thuộc vào các semiring nhất định có thể hạn chế phạm vi diễn giải, vì các semiring hoặc cấu trúc toán học khác cũng có thể mang lại những góc nhìn và phân tích có ý nghĩa. Phân tích này đòi hỏi kiến thức chuyên môn, có

thể gây cản trở đối với những người không chuyên, qua đó hạn chế tính dễ hiểu đối với con người.

3.8 Thảo luận bổ sung về giải thích dựa trên cơ chế attention

Cơ chế attention nổi bật trong các kỹ thuật giải thích tầm quan trọng của đặc trưng nội tại bởi các tranh cãi về chúng trong X-NLP. Mặc dù điểm số attention thường được sử dụng cho các giải thích nội tại và cục bộ nhằm định lượng tầm quan trọng của đặc trưng, các nghiên cứu lại đưa ra những quan điểm mâu thuẫn về giá trị diễn giải của nó [63,74]. Một số công trình cho rằng điểm số chú ý không cấu thành các giải thích thực sự [33], trong khi những nghiên cứu khác lại gợi ý rằng khi được áp dụng đúng cách, cơ chế attention có thể mang lại những hiểu biết ý nghĩa [79].

Một nghiên cứu đáng chú ý [28] đã chỉ ra rằng cơ chế attention đơn thuần có thể không phục vụ như những giải thích rõ ràng, trong khi độ nổi bật của cơ chế attention (tức là đạo hàm đầu ra đối với điểm attention) có thể cung cấp những hiểu biết thuyết phục hơn. Nghiên cứu cho thấy rằng mặc dù điểm attention không phân biệt được giữa các dự đoán khác nhau trong tác vụ Suy luận Ngôn ngữ Tự nhiên (NLI), độ nổi bật của attention lại tiết lộ những khác biệt đáng kể tương ứng với các nhãn dự đoán khác nhau, với các vùng quan trọng và có ý nghĩa được làm nổi bật.

Ngoài ra, để tăng cường độ tin cậy của điểm attention như những lời giải thích, [41] đã tinh chỉnh điểm attention bằng cách sử dụng tầm quan trọng của token từ các phương pháp giải thích thay thế trong một phương pháp tự huấn luyện. Điều này không chỉ củng cố mô hình tác vụ mà còn làm cho điểm attention phù hợp hơn với các kỹ thuật giải thích khác, cải thiện độ tin cậy và tính minh bạch của các giải thích dựa trên attention.

Điểm số attention cũng đối mặt với thách thức về độ tin cậy, có thể thay đổi dựa trên dữ liệu, tác vụ huấn luyện và thậm chí là sở thích cá nhân của người dùng. Trong lĩnh vực tài chính, chẳng hạn, các chủ doanh nghiệp khác nhau có thể tập trung vào những

khía cạnh khác nhau của cùng một bài báo, ảnh hưởng đến kết quả phân tích cảm xúc. Điều tích cực với người này có thể trở nên tiêu cực với người khác. Để giải quyết vấn đề này, một kỹ thuật Attention Query-driven Hierarchical (HQA) đã được giới thiệu [45], xem xét các biểu diễn nhúng của câu truy vấn cùng với các biểu diễn nhúng từ ngữ và câu để tạo ra các biểu diễn văn bản và điểm attention. Phương pháp này cho phép một bài báo tạo ra nhiều biểu diễn và ma trận attention tương ứng cho các truy vấn khác nhau, tăng tính linh hoạt và phù hợp của các giải thích dựa trên attention.

Chương 4: PHƯƠNG PHÁP THỰC HIỆN

4.1 Những vấn đề mở và ý tưởng giải quyết

4.1.1 Vấn đề Mở 1: Hạn chế của Phương pháp Nhiễu loạn Dữ liệu hiện tại

Các kỹ thuật hiện nay đánh giá tầm quan trọng của đặc trưng trong các mô hình XAI thường tạo ra tập dữ liệu nhiễu loạn từ dữ liệu gốc và quan sát sự thay đổi trong đầu ra của mô hình. Phương pháp này, lấy ý tưởng từ LIME, loại bỏ ngẫu nhiên một số đặc trưng trong dữ liệu gốc ban đầu. Việc nhiễu loạn này giúp khám phá hành vi cục bộ của mô hình xung quanh các đặc trưng bị nhiễu loạn. Sự thay đổi lớn trong đầu ra cho thấy mức độ quan trọng cao của đặc trưng bị nhiễu loạn, và ngược lại.

Tuy nhiên, việc ngẫu nhiên loại bỏ đặc trưng, chẳng hạn ở mức từ, có thể làm tổn thương cấu trúc ngữ pháp của câu. Khi đột ngột cắt bỏ một từ đóng vai trò ngữ pháp quan trọng, câu trở nên không hoàn chỉnh và gây ra sự thay đổi lớn trong dự đoán. Sự thay đổi này đôi khi không xuất phát từ sự vắng mặt của từ bị xóa mà đến từ sự phá vỡ ngữ pháp của câu. Do đó, sự thay đổi lớn trong dự đoán có thể dẫn đến ngộ nhận rằng từ bị xóa rất quan trọng, đánh giá quá cao tầm quan trọng của đặc trưng này.

Để khắc phục hạn chế trên, một số công trình đã cải tiến kỹ thuật nhiễu loạn bằng phương pháp masking, thay thế đặc trưng bằng token [MASK]. Điều này đảm bảo được sự hoàn chỉnh trong cấu trúc câu nhưng nội dung của từ ấy đã bị che khuất. Nhưng điểm yếu của phương pháp này là quên rằng mô hình SRL được fine-tune từ mô hình ngôn ngữ tiền huấn luyện trên tác vụ điền từ vào chỗ trống. Chính vì thế, mô hình tác vụ "thừa sức" đoán được từ bị che, vô tình khiến cho phương pháp này bị phụ thuộc vào mô hình tác vụ, ảnh hưởng đến tính model-agnostic. Điều đó dẫn đến thay đổi không đáng kể đến dự đoán của mô hình SRL kể cả khi từ bị che rất quan trọng, làm đánh giá thấp tầm quan trọng của đặc trưng này.

Nhận thấy được rủi ro tiềm ẩn từ các phương pháp nhiễu loạn trên, phương pháp Smart Substitution (SS) được khóa luận đề xuất để khắc phục được các hạn chế trên. Thay vì xóa hay masking đặc trưng từ thì khóa luận sẽ sử dụng phương pháp thay từ để tạo ra dữ liệu nhiễu loạn. Việc thay từ này phải đảm bảo hai yếu tố.

- Thứ nhất, từ được thay phải tương đồng về mặt ngữ pháp với từ gốc để đảm bảo cấu trúc ngữ pháp hoàn chỉnh và tương đương.
- Thứ hai, từ được thay phải đảm bảo đủ khác biệt về mặt ngữ nghĩa, tức từ mới phải trái nghĩa hoặc càng xa nghĩa với từ gốc càng tốt.

Hai yếu tố trên giúp bộ dữ liệu xáo trộn được tạo ra tương đồng về mặt ngữ pháp nhưng xa rời về mặt ngữ nghĩa với bộ dữ liệu gốc, giúp đánh giá hiệu quả sự biến động của dự đoán mô hình đến từ sự biến động về mặt ngữ nghĩa của từ mà ta cần đo lường tầm quan trọng. Hơn nữa, để tránh việc làm nhiễu loạn đến kết quả bởi số lượng lớn các mạo từ, giới từ không mang quá nhiều ngữ nghĩa trong tập dữ liệu, phương pháp chỉ tập trung vào các thực từ có vai trò ngữ nghĩa quan trọng trong câu như động từ chính, danh từ, tính từ và các trạng từ.

Minh họa về phương pháp *Smart Substitution*:

Câu gốc : “He felt in love with a young **lady** that he met at the party.”

Thay thế thông thường: “He felt in love with a young **woman** that he met at the party.”

Smart Substitution : “He felt in love with a young **man** that he met at the party.”

“lady” và “man” đều là danh từ và có nghĩa trái ngược nhau hoàn toàn, thỏa mãn các yếu tố của phương pháp Smart Substitution. Trong khi “lady” và “woman” cùng loại từ nhưng không khác biệt về ngữ nghĩa.

Phương pháp Smart Substitution là một phương pháp giải thích hậu nghiệm, không can thiệp vào tham số cũng như cấu trúc nội tại của mô hình, mà sử dụng dữ liệu đầu vào và kết quả dự đoán để đưa ra các giải thích. Tính model-agnostic của phương pháp

cho phép áp dụng cho mọi loại mô hình, từ mạng nơ-ron đơn giản cho đến các mô hình Transformer, mà không cần thay đổi quá trình huấn luyện ban đầu. Phương pháp này linh hoạt, khả năng ứng dụng cao, giúp các nhà phát triển hiểu rõ hơn về mô hình, cải thiện tính minh bạch cũng như độ tin cậy của các hệ thống NLP trong thực tế. Bên cạnh đó, kết quả thực nghiệm của khóa luận (**Chương 5**) cho thấy phương pháp này đem đến lời giải thích có tính trung thực hơn các phương pháp làm nhiễu loạn dữ liệu truyền thống, đảm bảo rằng các giải thích phản ánh chính xác cách mô hình hoạt động trong thực tế. Đồng thời, phương pháp cung cấp các giải thích dễ hiểu và có ý nghĩa đối với người dùng, tạo điều kiện cho việc tương tác và ra quyết định. Chi tiết được làm rõ trong **Mục 4.2**.

Dưới đây là quy trình tổng quát của phương pháp đề xuất Smart Substitution:

- Bước 1: Xây dựng bộ dữ liệu đánh dấu vị trí và loại từ của các thực từ trong câu.
- Bước 2: Xây dựng danh sách ứng viên cho từng loại thực từ.
- Bước 3: Chọn từ thay thế theo phương pháp Smart Substitution.

4.1.2 Vấn đề Mở 2: Khai phá Thông tin trong Không gian ẩn

Không gian ẩn (vector biểu diễn) trong các mô hình ngôn ngữ mang lại một sự tiến bộ vượt bậc trong Xử lý Ngôn ngữ Tự nhiên (NLP) nhờ khả năng nắm bắt tri thức ngữ nghĩa và cú pháp của từ ngữ. Các vector này mã hóa thông tin ngữ nghĩa bằng cách biểu diễn từ trong không gian nhiều chiều, nơi các từ có nghĩa tương tự được đặt gần nhau. Kỹ thuật như Word2Vec, GloVe và BERT đã cải thiện đáng kể khả năng của các mô hình ngôn ngữ trong việc hiểu và xử lý ngôn ngữ tự nhiên. Chúng học được các quan hệ ngữ nghĩa phức tạp, nhận diện ngữ cảnh của từ trong câu, và thực hiện các tác vụ NLP như phân loại văn bản, dịch máy, và trả lời câu hỏi một cách hiệu quả.

Một trong những điểm mạnh của không gian latent là khả năng học và biểu diễn các quan hệ ngữ nghĩa phức tạp giữa các từ, chẳng hạn như mối quan hệ giữa từ "vua" và

"nữ hoàng" hoặc "Paris" và "Pháp". Khả năng này giúp các mô hình không chỉ hiểu nghĩa của từ mà còn hiểu được mối quan hệ giữa chúng. Điều này đặc biệt quan trọng trong các tác vụ đòi hỏi sự hiểu biết sâu sắc về ngữ nghĩa như dịch máy, nơi mà việc dịch đúng đòi hỏi phải xác định rõ mối quan hệ ngữ nghĩa giữa các từ và cụm từ trong ngữ cảnh.

Ngoài ra, các mô hình như BERT không chỉ biểu diễn từ ngữ mà còn biểu diễn toàn bộ câu hoặc đoạn văn, cho phép chúng nắm bắt được ngữ cảnh toàn cục. Điều này làm tăng khả năng hiểu biết ngữ cảnh của mô hình, giúp nó đưa ra các dự đoán chính xác hơn trong nhiều tác vụ khác nhau, từ phân loại văn bản đến Nhận diện thực thể có tên (NER).

Tri thức ngữ nghĩa này không chỉ giúp các mô hình ngôn ngữ trở nên mạnh mẽ và linh hoạt hơn, mà còn mở ra nhiều khả năng ứng dụng thực tiễn trong nhiều lĩnh vực khác nhau, từ y học (phân tích văn bản y khoa, chẩn đoán bệnh tật) đến kinh doanh (phân tích cảm xúc, chatbot). Không gian ẩn thực sự đã thay đổi cách con người tiếp cận và giải quyết các vấn đề liên quan đến ngôn ngữ tự nhiên, làm cho các mô hình trở nên thông minh và hữu dụng hơn.

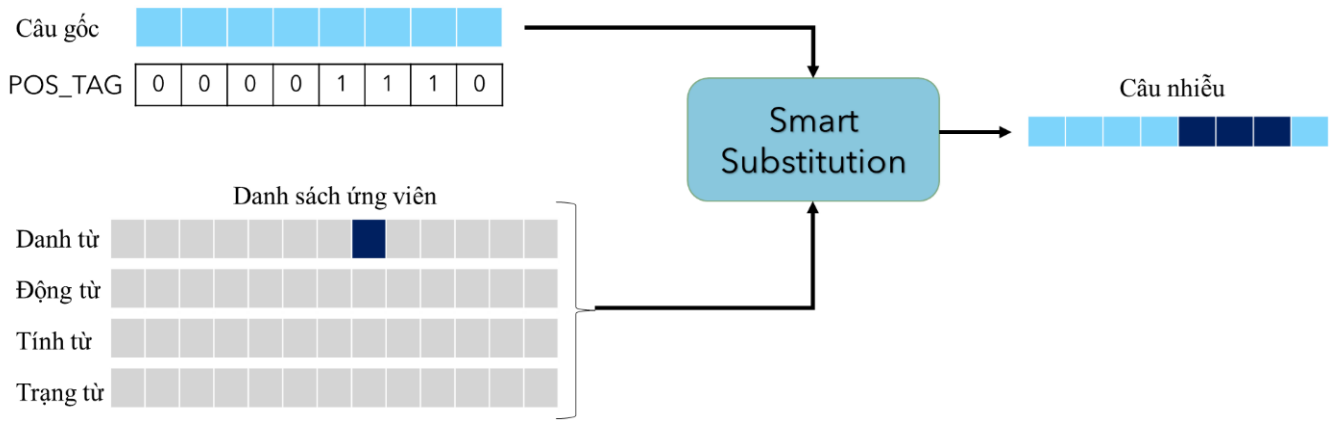
Dựa trên sức mạnh này của không gian latent, khóa luận đề xuất nghiên cứu để kiểm tra xem vector biểu diễn của các từ trong mô hình ngôn ngữ có thể mã hóa tầm quan trọng của từ trong tác vụ SRL hay không. Để thực hiện quá trình này, đầu tiên tri thức về PAS sẽ được tách biệt ra khỏi tri thức của vector biểu diễn và so sánh với độ quan trọng của từ trong dự đoán SRL. Nếu chúng có mối tương quan cao với nhau, điều đó cho thấy vector biểu diễn từ có mã hóa tầm quan trọng của từ ấy trong tác vụ gán nhãn ngữ nghĩa, và ngược lại.

Quá trình finetuning điều chỉnh tối ưu hóa hệ thống tham số nhằm tập trung vào tác vụ SRL, dạy cho mô hình học thêm tri thức về PAS. Sự khác biệt giữa vector biểu diễn trước và sau khi finetune sẽ phản ánh lượng tri thức về PAS cần tách lọc. Do đó, khóa luận sẽ finetune mô hình BioBert trên tác vụ SRL và tính toán tri thức về PAS. Sau đó,

hệ số tương quan Spearman được sử dụng để ước lượng mối quan hệ giữa tri thức về PAS và tầm quan trọng của từ đối với dự đoán. Chi tiết cụ thể được trình bày ở **Mục 4.3**.

4.2 Giải thích tầm quan trọng của đặc trưng từ bằng Smart Substitution

Dữ liệu đầu vào của phương pháp này bao gồm dữ liệu được đánh dấu các vị trí từ cần đánh giá tầm quan trọng và danh sách từ ứng viên tương ứng cho mỗi vị trí ấy. Sau đó, phương pháp Smart Substitution sẽ lựa chọn từ thay thế tốt nhất trong các ứng viên để thay vào dữ liệu gốc, từ đó khóa luận sẽ có được bộ dữ liệu xáo trộn theo phương pháp Smart Substitution đề xuất. Qua đó, khóa luận có thể đánh giá được hành vi của mô hình dựa trên bộ dữ liệu mới này và định lượng được tầm quan trọng của đặc trưng dựa vào các đại lượng đo tầm quan trọng của đặc trưng. Quy trình tổng quát của phương pháp được mô tả trong ***Hình 4***.



Hình 4. *Mô tả khái quát quá trình Smart Substitution.*

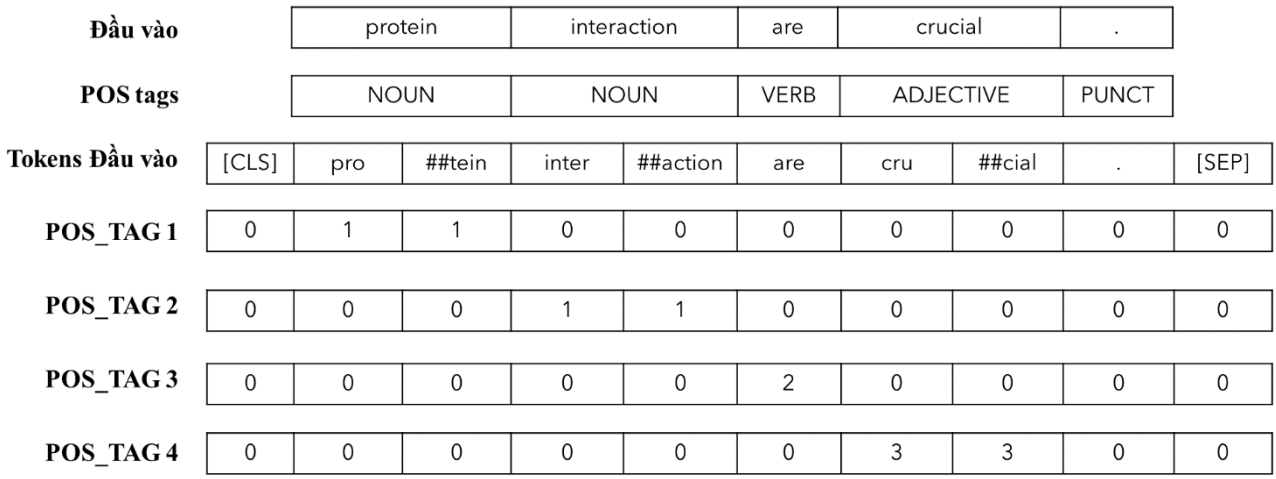
Smart Substitution là một phương pháp giải thích hậu nghiệm, hoàn toàn chỉ dựa trên dữ liệu đầu vào và dự đoán của mô hình để đưa ra lời giải thích. Do đó, bộ dữ liệu từ kho văn bản cần được thông qua các bước xử lý, biến đổi để phù hợp với đầu vào của phương pháp giải thích. Đầu tiên, các thực từ có vai trò ngữ nghĩa quan trọng trong câu cần được đánh dấu, xử lý và lưu trữ để khai thác các thông tin quan trọng về loại từ, vị

trí token cũng như vector biểu diễn của chúng. Sau đó, để phương pháp có thể lựa chọn được từ ứng viên phù hợp với yếu tố về loại từ, khóa luận cần xây dựng một danh sách ứng viên cho từng vị trí cần thay thế. Các câu gốc cùng thực từ được đánh dấu và danh sách ứng viên sẽ được đi qua phương pháp Smart Substitution tính toán, xử lý và trả ra kết quả với bộ dữ liệu mới đảm bảo dữ liệu mới đủ khác biệt với dữ liệu gốc để đánh giá tầm quan trọng đặc trưng từ một cách trung thực nhất. Sau đây là quá trình thực hiện chi tiết:

Bước 1: Xây dựng bộ dữ liệu đánh dấu vị trí và loại từ.

Đối với mỗi bộ nhúng của một câu văn bản đầu vào, khóa luận thực hiện gán vai trò ngữ pháp cho mỗi từ trong câu như danh từ, động từ, tính từ và trạng từ bằng thư viện SpaCy. Để dễ dàng đánh giá riêng lẻ từng thực từ trong câu, mỗi thực từ sẽ có một vector POS_TAG riêng. Như vậy, với mỗi câu gốc, khóa luận sẽ có số vector POS_TAG khác nhau tùy thuộc vào số lượng thực từ trong câu. Vector POS_TAG vừa giúp đánh dấu vị trí các token của thực từ, đồng thời cho biết vai trò ngữ pháp của thực từ ấy, được minh họa trong ***Hình 5.***

POS_TAG là một vector one-hot có độ dài bằng vector mã hóa của câu, được khởi tạo với tất cả giá trị 0 và tại các vị trí token là thực từ sẽ được đánh dấu theo pos_tag_mapping = {1: danh từ, 2: động từ, 3:tính từ, 4: trạng từ}.



Hình 5. *Minh họa cách vector POS_TAG được tạo ra từ một câu gốc.*

Như vậy, với mỗi mẫu trong bộ dữ liệu tạo ra sẽ gồm ID, vector nhúng câu gốc, masked_sentence (che giấu các vị trí token của thực từ bằng token [MASK]), vector POS_TAG, vector nhúng của thực từ bị che và nhãn "vàng" của các đối số cho tác vụ SRL. Ở đây vector nhúng của thực từ bị che được giữ lại để phục vụ cho việc xây dựng danh sách ứng viên được mô tả ở *Bước 2*.

Bước 2: Xây dựng danh sách ứng viên cho từng loại thực từ.

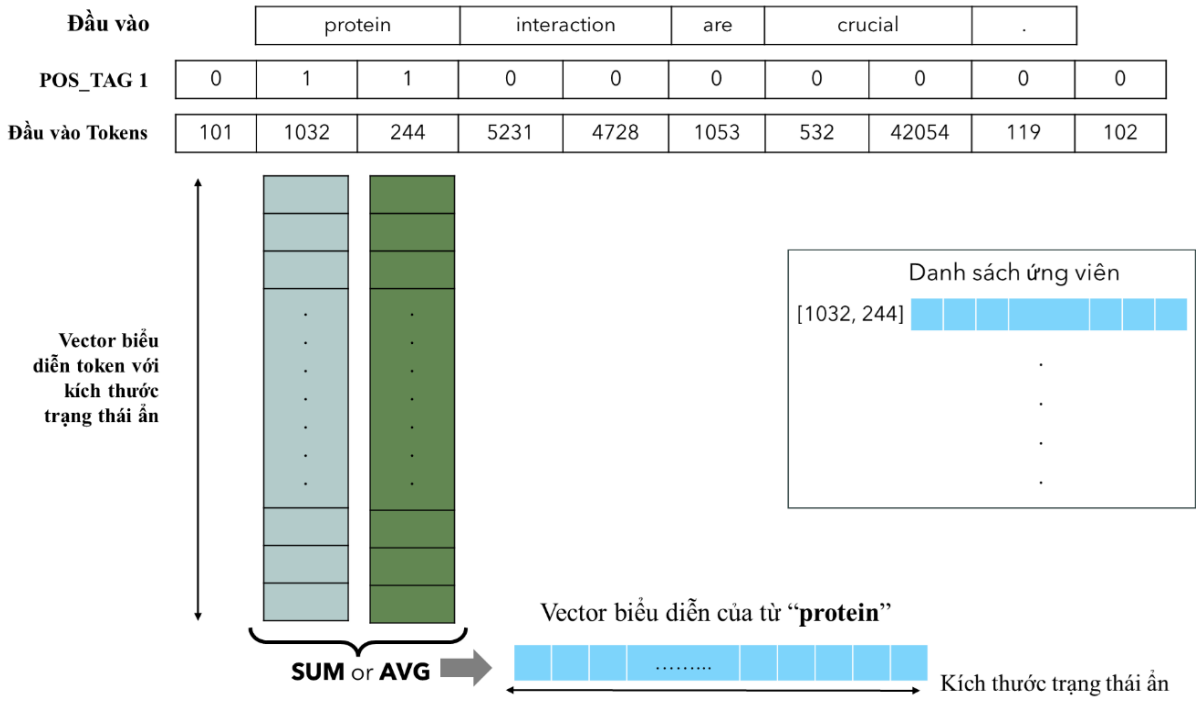
Danh sách ứng viên này sẽ giúp lựa chọn được loại thực từ phù hợp với vị trí của từ cần đánh giá độ quan trọng, đồng thời vừa giúp giới hạn lại không gian của từ ứng viên trong khuôn khổ của ngữ cảnh huấn luyện. Vì nếu không xây dựng danh sách này sẽ có vô số từ ứng viên có vai trò ngữ pháp tương đương với từ cần đánh giá và sẽ rất khó kiểm soát.

Dựa vào thông tin về vị trí và loại từ mà vector POS_TAG cung cấp, những thực từ có cùng vai trò ngữ pháp sẽ được tách riêng và lưu trữ vào cùng một danh sách tương ứng với loại từ của vai trò ngữ pháp ấy. Bên cạnh danh sách lưu trữ các từ có cùng vai trò ngữ pháp, các vector biểu diễn của từ ấy cũng được lưu trữ thành cặp với vector từ

trong danh sách. Do đó, bộ dữ liệu được mã hóa ở *Bước 1* được đưa vào mô hình ngôn ngữ để lấy ra vector biểu diễn cho từng thực từ.

Như vậy, khóa luận sẽ có được 4 danh sách ứng viên cho 4 loại thực từ, bao gồm vector nhúng và vector biểu diễn của từ ấy. Mục đích của việc giữ lại vector nhúng của từ trong danh sách ứng viên là dùng để thay thế vào vị trí của từ gốc ban đầu trong câu. Ở đây, vector biểu diễn của từ sẽ là SUM hoặc AVG tất cả các vector biểu diễn của token tạo nên từ ấy. Vì vốn dĩ bộ nhúng BioBert sử dụng phương pháp WordPiece để xử lý văn bản, do đó một từ có thể được mã hóa thành nhiều token. ***Hình 6*** minh họa cho quá trình hình thành danh sách ứng viên.

Vector biểu diễn cho mỗi câu có kích thước (85, 768), với 85 là kích thước của câu mã hóa, tức số lượng token trong câu và 768 là kích thước tầng ẩn (hidden layer size) của mô hình, chính là kích thước vector biểu diễn mỗi token trong câu.



Hình 6. *Minh họa quá trình lấy vector biểu diễu từ mô hình và xây dựng danh sách ứng viên.*

Sau khi đã đánh dấu được vị trí từ cần thay thế trong câu cũng như xây dựng danh sách từ ứng viên, khóa luận tiến hành lựa chọn từ phù hợp trong danh sách ứng viên ứng với từng vị trí cần thay thế với yếu tố không liên quan về ngữ nghĩa. Ở đây, các độ tương đồng giữa các vector biểu diễn từ sẽ được sử dụng để thỏa mãn yếu tố từ thay thế không liên quan về ngữ nghĩa với từ gốc cần đánh giá.

Bước 3: Lựa chọn từ thay thế theo phương pháp Smart Substitution

Smart Substitution chọn từ thay thế dựa trên hai yếu tố cốt lõi: cùng loại từ nhưng khác biệt ngữ nghĩa. Đối với yếu tố cùng loại từ, từ thay thế sẽ được lựa chọn từ danh sách ứng viên cùng loại. Để đảm bảo khác biệt về ngữ nghĩa, khóa luận sử dụng độ tương đồng cosine để kiểm tra khoảng cách ngữ nghĩa của từ gốc và các từ ứng viên. Từ ứng viên được chọn trái nghĩa với từ gốc (tương ứng với độ tương đồng gần -1) hoặc không liên quan về ngữ nghĩa với từ cần thay thế (tương ứng độ tương đồng gần 0). Công thức độ tương đồng cosine như sau:

$$cosine_{similarity}(v_1, v_2) = \frac{v_1.v_2}{|v_1|.|v_2|} \in [-1;\ 1] \quad (18)$$

Trong đó, v_1, v_2 là vector biểu diễn của từ gốc và từ ứng viên ($v_1, v_2 \in \mathbb{R}^{85\times768}$).

Độ tương đồng cosine tồn tại một hạn chế lớn là chỉ dựa vào góc quay của hai vector mà bỏ qua hoàn toàn độ dài của chúng. Điều này có nghĩa là hai vector có thể trùng nhau về hướng nhưng có độ dài khác nhau vẫn sẽ có độ tương đồng cosine bằng 1. Ví dụ, một vector dài và một vector ngắn nhưng cùng hướng sẽ có độ tương đồng cosine tối đa, mặc dù trong thực tế, sự tương quan giữa chúng có thể không chặt chẽ. Điều này cho thấy độ đo cosine không hoàn toàn đánh giá khách quan và chính xác mối tương quan giữa hai vector, vì nó không xem xét đến sự khác biệt về độ dài giữa các vector.

Để khắc phục hạn chế này, khóa luận đề xuất một độ đo mới, hiệu quả hơn, được cải tiến từ độ đo cosine, gọi là độ tương đồng cosine_module. Độ tương đồng cosine_module kết hợp cả độ tương đồng cosine và độ tương đồng về độ dài (module)

giữa hai vector. Cụ thể, độ tương đồng cosine_module là tích vô hướng giữa độ tương đồng cosine và một yếu tố điều chỉnh dựa trên độ dài. Yếu tố điều chỉnh này được tính bằng 1 trừ đi tỉ số độ khác biệt độ dài trên tổng độ dài hai vector. Công thức này cho phép đo lường không chỉ sự tương đồng về hướng mà còn sự tương đồng về độ dài, mang lại cái nhìn toàn diện hơn về mối tương quan giữa hai vector. Công thức được trình bày như sau:

$$cosine_{module} = \underbrace{(1 - \delta)}_{\textbf{Module}} . \underbrace{cosine\_similarity(v_1, v_2)}_{\textbf{Cosine}} \quad (19)$$

Trong đó, $\delta = \frac{||v_1|-|v_2||}{|v_1|+|v_2|}$ là tỉ số độ khác biệt độ dài trên tổng độ dài hai vector.

Như vậy, đối với vector biểu diễn từ, khóa luận sẽ có 2 phương thức biểu diễn là lấy tổng hoặc trung bình (tương ứng SUM hoặc AVG) của tất cả vector biểu diễn token tạo nên từ ấy. Còn đối với độ tương đồng về ngữ nghĩa của hai vector biểu diễn, khóa luận sử dụng cosine và cosine_module. Để lựa chọn ra từ thay thế thỏa mãn yếu tố không liên quan về ngữ nghĩa, khóa luận có 2 lựa chọn, là thay thế từ trái nghĩa (tương ứng similarity gần -1) và thay thế từ không liên quan đến nghĩa của từ gốc (tương ứng similarity gần 0). Tổ hợp các khía cạnh trên, khóa luận sẽ có tổng cộng 8 phương pháp lựa chọn từ thay thế. Kết quả phương pháp nào hiệu quả hơn sẽ được chứng minh bằng thực nghiệm trong chương 5.

Bảng 2 là thuật toán từ một câu gốc tạo thành 8 câu mới theo 8 phương pháp thay thế:

***Bảng 2**. Thuật toán Tạo nhiều câu nhiễu từ một câu gốc.*

Thuật toán. **Tạo nhiều câu nhiễu từ một câu gốc**

1: **Function: GENERATE_NEW_SENTENCES (D, Dict_candidate)**
Đầu vào:
- *D* = DataFrame lưu thông tin của câu dữ liệu gốc gồm các thuộc tính: ID, vector POS_TAG, sum_vector, avg_vector (sum_vector và avg_vector là vector biểu diễn từ cần thay thế trong câu).
- *Dict_candidate* = Dictionary của danh sách ứng viên với key là loại từ và value là Danh sách từ tương ứng.

2: *L_word* = []; // Khởi tạo List trống để lưu giữ Danh sách ứng viên phù hợp.
3: // Lấy ra vị trí của các token cần thay thế chính là vị trí các giá trị khác 0 trong vào vector POS_TAG.
4: *masked_index* = (POS_TAG != 0);
5: // Lấy ra giá trị đầu tiên của masked_index chính là giá trị của loại từ cần thay thế.
6: *type_word = masked_index*[0];
7: // Kiểm tra loại từ của từ cần thay thế và đưa vào danh sách ứng viên tương ứng.
8: **if** *type_word* == 1 **then**: *L_word = Dict_candidate*[‘noun’];
9: **elif** *type_word* == 2 **then**: *L_word = Dict_candidate*[‘nerb’];
10: **elif** *type_word* == 3 **then**: *L_word = Dict_candidate*[‘adj’];
11: **elif** *type_word* == 4 **then**: *L_word = Dict_candidate*[‘adv’];
12: **end if**
13: // Đưa danh sách ứng viên vừa tìm được và các vector biểu diễn của từ gốc vào hàm Pick Candidate Words để tìm từ ứng viên thay thế cho từng phương pháp.
14: *list_new_words* = **Pick Candidate Words**(*sum_vector, avg_vector, L_word*);
15: *list_new_sentences* = []; // Khởi tạo danh sách để lưu trữ các câu mới.
16: // Thay thế từng từ vào câu gốc để tạo ra các câu mới.
17: **forall** *new_word* ∈ *list_new_words* **do**
18: *new_sentence* = replace (*origin_id, new_word, origin_word*);
19: **end for**
20: **return** *new_sentences*
21: **end function**

Trong thuật toán trên, $type\_word$ giúp tìm ra từ loại của từ cần thay thế, và tương ứng với từng type_word sẽ có Danh sách từ ứng viên L_word tương ứng. Sau đó, vector biểu diễn của từ cần thay thế được biểu diễn dưới dạng sum_vector và avg_vector cùng với Danh sách L_word được xây dựng trước đó đi qua chức năng **Pick Candidate Words** để tìm ra ứng viên thích hợp nhất cho từng phương pháp thay thế. Như vậy, mỗi thực từ trong câu sẽ có 8 từ thay thế mới tương ứng với 8 phương pháp, tạo ra 8 câu dữ liệu mới được làm nhiễu loạn từ một câu văn bản gốc.

Trong đó, **Pick Candidate Words** là thuật toán tìm ra từ ứng viên phù hợp cho từng phương thức $Method$ truyền vào (Sum hoặc Avg). Đầu vào thuật toán cũng cần có vector biểu diễn từ gốc và danh sách các vector biểu diễn từ ứng viên L_word để tính toán độ tương đồng và lựa chọn ứng viên phù hợp thỏa mãn các yếu tố của phương pháp Thay thế Thông minh. Thuật toán này có nhiệm vụ tìm ra ứng viên có khoảng cách ngữ nghĩa xa nhất với từ gốc (tương ứng độ tương đồng xấp xỉ -1) và ngữ nghĩa không liên quan đến từ gốc (tương ứng độ tương đồng gần 0). Đầu tiên, khóa luận tính toán một danh sách giá trị độ tương đồng L_result cho tất cả các ứng viên trong danh sách. Ở đây, L_result sẽ tính toán cả giá trị cosine và cosine module cho từng cặp vector biểu diễn từ. Do đó, mỗi phần tử trong L_result sẽ bao gồm cos_val, cosmo_val, cos_word, cosmo_word tương ứng giá trị cosine, cosine module và các từ có giá trị cosine, cosine module tương ứng. Với độ đo cosine, khóa luận sắp xếp L_result theo giá trị đầu với thứ tự tăng dần (dòng 9) và để chọn được ứng viên có giá trị cosine gần -1, khóa luận chỉ cần lấy ra phần tử đầu tiên trong mảng được sắp xếp $Sorted\_cos$ (dòng 11). Tương tự với cosine module, L_result sẽ sắp xếp theo giá trị thứ hai trong mảng (dòng 13) và phần tử đầu trong mảng sắp xếp được lấy ra tương ứng với cosine module gần -1 (dòng 14).

Đối với độ tương đồng gần 0 (dòng 16 → 21), khóa luận cũng tiến hành sắp xếp chuỗi giá trị độ tương đồng theo thứ tự tăng dần và lấy ra phần tử đầu trong chuỗi, tuy

nhiên chuỗi sẽ được sắp xếp theo giá trị tuyệt đối, khác với độ tương đồng gần -1 là lấy giá trị trực tiếp. Thuật toán **Pick Candidate Words** được trình bày trong ***Bảng 3***.

Bảng 3. *Thuật toán Chọn từ ứng viên cho từng phương thức.*

Thuật toán: Chọn từ ứng viên cho từng phương thức

1: **Function: PICK_CANDIDATE_WORDS (V, Method, L_word)**
Đầu vào:
- *V*: Vector biểu diễn từ gốc trong câu.
- *Method*: SUM hoặc AVG các vector biểu diễn tokens cấu tạo nên từ ứng viên.
- *L_word*: Danh sách từ ứng viên bao gồm vector nhúng từ, sum_vector, avg_vector.

2: *L_result* = [] // Khởi tạo List trống để lưu trữ giá trị độ tương đồng.
3: // Duyệt qua từng phần tử trong Danh sách *L_word* để tính độ tương đồng cosine, cosine module với vector biểu diễn từ gốc *V* và lưu vào *L_result.*
4: **forall** *item* in *L_word* **do**
5: *L_result* append **cal_similarity**(*V, Method, item*);
6: **end for**
7: // Sắp xếp mảng giá trị *L_result* để lấy ra phần tử nhỏ nhất (cosine gần -1).
8: *sorted_cos* = sort(*L_result,* key = first element); // bằng phần tử thứ nhất theo thứ tự tăng dần.
9: // Ứng viên được chọn có độ tương đồng gần -1
10: *neg_cos_val, neg_cos_word* = *sorted_cos*[0]; // phần tử nhỏ nhất trong *sorted_cos*
11: *sorted_cosmo* = sort (*L_result*, key = second element) // Tương tự cho cosine module được sort bằng phần tử thứ hai theo thứ tự tăng dần.
12: neg_cosmo_val, neg_cosmo_word = sorted_cosmo[0];
13: // Sắp xếp mảng kết quả *L_result* bằng giá trị tuyệt đối của độ tương đồng để lấy ra phần tử có độ tương đồng gần 0 nhất.
14: *sorted_abs_cos* = sort (*L_result,* key = |first element|) // sort bằng phần tử thứ nhất theo thứ tự tăng dần.
15: // Chọn ứng viên có độ tương đồng gần 0 là phần tử nhỏ nhất trong mảng sắp xếp *sorted_abs_cos*.
16: *pos_cos_val, pos_cos_word = sorted_abs_cos*[0];
17: // Tương tự cho cosine module gần 0.
18: *sorted_abs_cosmo* = sort(*L_result,* key = |second element|) //sort bằng phần tử thứ hai theo thứ tự tăng dần.
19: *pos_cosmo_val, pos_cosmo_word = sorted_abs_cosmo*[0];
20: // Kết quả 4 phương thức (cos, cosmo) x (gần -1, gần 0) cho mỗi *Method* truyền vào.
21: **return** { “neg_cos”: (neg_cos_val, neg_cos_word),
“neg_cosmo”: (neg_cosmo_val, neg_cosmo_word),

"pos_cos": (pos_cos_val, pos_cos_word),
"pos_cosmo": (pos_cosmo_val, pos_cosmo_word)}
22: **end function**

4.3 Kiểm tra sự hiện hữu tầm quan trọng đặc trưng trong vector biểu diễn

Để có thể kiểm tra được có hay không sự tồn tại về tầm quan trọng của đặc trưng trong không gian latent của mô hình. Khóa luận sẽ thực hiện phân tích, đánh giá tầm quan trọng của đặc trưng trong không gian ấy. Đồng thời phân tách tri thức trong vector biểu diễn từ để từ đó tìm ra mối tương quan giữa tầm quan trọng của đặc trưng với khối tri thức mà vector biểu diễn chứa đựng. Mối tương quan cao (ngưỡng 0.5) cho thấy tri thức được mã hóa trong vector biểu diễn có tác động đến tầm quan trọng của từ và ngược lại.

Chi tiết về quá trình kiểm tra tầm quan trọng đặc trưng trong không gian ẩn:

Bước 1: Xây dựng chuỗi giá trị độ quan trọng của từ.

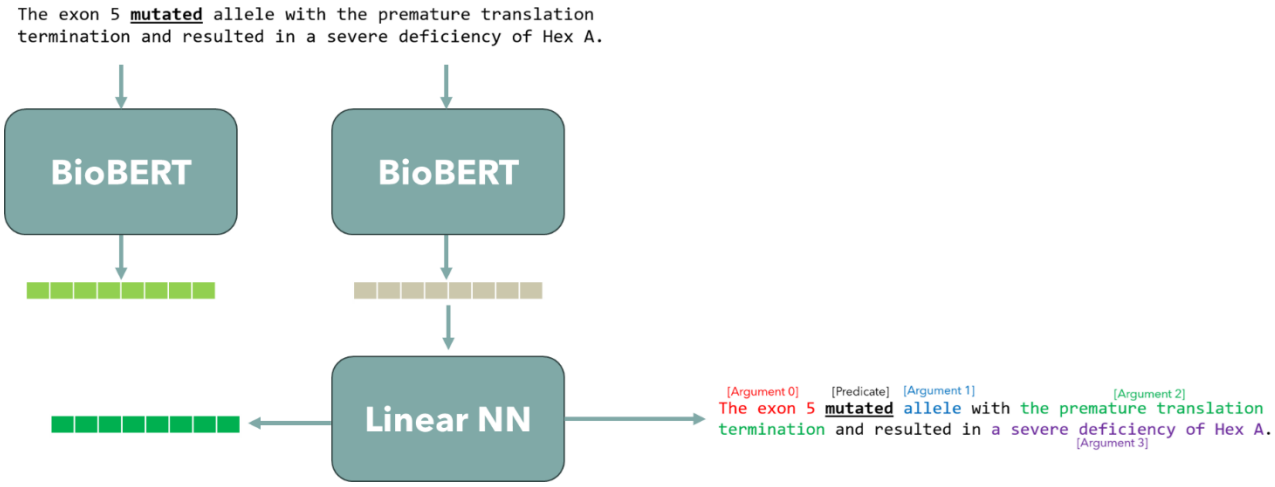
Chuỗi giá trị này được hình thành bằng cách ghép tất cả giá trị *influence* của từng từ (*influence* là phần trăm đóng góp của một từ vào dự đoán đầu ra của mô hình, và sẽ được trình bày chi tiết trong **Chương 5**), được tính toán dựa trên phương pháp chọn từ thay thế tốt nhất (đề cập ở **Mục 4.2**), tạo thành một chuỗi$\{inf\}$. Thuật toán tính toán các giá trị tầm quan trọng của thực từ trong bộ dữ liệu được thể hiện trong ***Bảng 4***.

Đầu vào của thuật toán bao gồm bộ dữ liệu D_origin gồm các câu gốc, dự đoán của mô hình, giá trị vector xác suất cho bộ dữ liệu gốc và $D\_perturb$ gồm các câu dữ liệu xáo trộn, dự đoán của mô hình và giá trị vector xác suất cho bộ dữ liệu xáo trộn ấy. Sau đó, khóa luận tiến hành tính toán điểm *influence, relevance, Brier score* cho từng mẫu dữ liệu xáo trộn, dòng 13 đến 17 (được mô tả chi tiết trong phần thực nghiệm **Chương 5**). Sau đó, thuật toán khởi tạo một $Dictionary\ D$ với *key* là đối số của từng động từ

trong bộ dữ liệu và *value* là các giá trị định lượng sức ảnh hưởng và độ hữu ích của từng từ tác động đến đối số đó (dòng 21 đến 24). Ngoài ra thuật toán còn khởi tạo L_score để lưu trữ độ quan trọng của từ mà không cần gom nhóm theo đối số của mỗi nhóm động từ như D đã làm, mục đích để biểu diễn độ tương quan với chuỗi giá trị định lượng tri thức PAS được miêu tả trong *Bước 2*.

Bước 2: Hình thành chuỗi giá trị định lượng tri thức PAS.

Tri thức về PAS được mã hóa trong không gian latent của dữ liệu, ước lượng bằng cách đo sự chênh lệch giữa vector biểu diễn từ trước và sau khi tinh chỉnh (finetuning). Vector biểu diễn được lấy từ lớp cuối cùng (last_hidden_state) của mô hình. Bởi vì last_hidden_state là kết quả của quá trình tiền huấn luyện của mô hình trên một kho dữ liệu lớn để học các đặc trưng ngữ cảnh, do đó vector này chứa đựng thông tin rất phong phú và có giá trị về ngữ nghĩa và ngữ cảnh của các từ xung quanh. Quá trình ước lượng tri thức PAS được minh họa trong ***Hình 7.***



Hình 7. *Hình minh họa quá trình rút trích tri thức PAS từ vector biểu diễn của mô hình trước và sau khi finetuning (tương ứng vector xanh nhạt và xanh đậm).*

Khóa luận đưa ra hai giả thuyết về mối quan hệ giữa hai vector này:

Giả thuyết 1: Giả định rằng các từ quan trọng trong SRL, tự thân nó đã mang rất nhiều tri thức về PAS, cho nên sau quá trình tinh chỉnh nó sẽ không hấp thụ được thêm bao nhiêu về tri thức PAS nữa, dẫn đến sự thay đổi ít. Còn đối với những từ ít quan trọng, sự khác biệt trong vector biểu diễn của chúng lớn do được tiếp nhận thêm thông tin về PAS. Tức độ tương đồng giữa vector biểu diễn từ trước và sau khi finetune càng lớn thì độ quan trọng của từ đối với dự đoán SRL càng cao.

Như vậy, đối với *Giả thuyết* 1, độ tương đồng sẽ được dùng để ước lượng tri thức về PAS. Chuỗi giá trị độ tương đồng ($\hbar_1$) được hình thành bằng cách ghép toàn bộ giá trị độ tương đồng của thực từ trong bộ dữ liệu. Độ tương đồng được nhắc đến ở 3 hình thức là cosine, cosine_module và -|element-wise substraction|, tương ứng sẽ có $\hbar_{1a}, \hbar_{1b}, \hbar_{1c}$.

Giả thuyết 2: Với những từ quan trọng, quá trình finetuning sẽ giúp vector biểu diễn tập trung tích lũy rất nhiều tri thức về PAS; trong khi những từ ít ảnh hưởng đến dự đoán thì finetuning sẽ "phớt lờ" những từ ấy, cho nên tri thức của tác vụ không tích lũy được bao nhiêu, dẫn đến hai vector có khác biệt rất lớn. Do đó, đưa đến giả thuyết độ khác biệt trong hai vector biểu diễn từ trước và sau khi tinh chỉnh càng lớn cho thấy sức ảnh hưởng đến dự đoán của tác vụ càng nhiều.

Như vậy trong giả thuyết này, chuỗi giá trị định lượng tri thức PAS được sử dụng sẽ là độ khác biệt. Chuỗi giá trị độ khác biệt ($\hbar_2$) được hình thành bằng cách ghép tất cả giá trị khác biệt của các thực từ trong bộ dữ liệu. Ở đây, độ khác biệt được dùng ở 3 hình thức là (1 – cosine), (1 – cosine_module) và |(element-wise substraction)|, tương ứng sẽ có $\hbar_{2a}, \hbar_{2b}, \hbar_{2c}$.

Sau khi xây dựng được chuỗi các giá trị ước lượng tri thức về PAS, các chuỗi này sẽ được biểu diễn cùng với chuỗi giá trị độ quan trọng của từ để tìm ra mối tương quan giữa chúng. Việc biểu diễn này giúp chúng ta có cái nhìn sâu sắc hơn về cách mà các yếu tố khác nhau trong PAS ảnh hưởng đến tầm quan trọng của từ trong việc dự đoán SRL.

Bước 3: Biểu diễn độ tương quan giữa tri thức PAS và độ quan trọng của từ đối với dự đoán SRL.

Để tiến hành bước này, khóa luận sử dụng hệ số tương quan Spearman. Hệ số tương quan Spearman là một phương pháp thống kê để đo lường mối quan hệ giữa hai biến số dựa trên thứ hạng của chúng. Công thức hệ số Spearman như sau:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (20)$$

r_s : Hệ số tương quan Spearman.

d_i: Hiệu giữa hai thứ hạng của cặp giá trị thứ i của hai vector.

n : Tổng số cặp giá trị. Trong trường hợp này, n là tổng số thực từ.

Spearman đo lường mối quan hệ đơn điệu giữa hai biến, nghĩa là nó không chỉ đo lường mối quan hệ tuyến tính mà còn có thể bắt được các mối quan hệ phi tuyến tính. Điều này làm cho hệ số Spearman đặc biệt hữu ích khi làm việc với các dữ liệu thực nghiệm có thể chứa các giá trị ngoại lai hoặc không tuân theo phân phối chuẩn. Chẳng hạn xét các ví dụ sau để thấy sự khác biệt rõ rệt giữa hệ số tương quan Spearman và hệ số tương quan Pearson.

A = [1, 2, 3, 6, 100]

B = [1, 3, 4, 6, 7]

Hệ số Pearson r(A, B) = $\frac{n\sum A_iB_i - \sum A_i.\sum B_i}{\sqrt{[n\sum A_i^2 - (\sum A_i)^2][n\sum B_i^2 - (\sum B_i)^2]}} \approx 0.686$

Hệ số tương quan Spearman = 1 (áp dụng công thức (20))

Hoặc xét tập dữ liệu sau:

X = [1, 2, 3, 4, 5]

Y = [1, 4, 9, 16, 25] ($Y=X^2$)

Hệ số Pearson r(X, Y) ≈ 0.981

Hệ số tương quan Spearman = 1

Từ các ví dụ trên cho thấy Pearson nhạy cảm với các giá trị biệt lệ và không phản ảnh chính xác mối quan hệ phi tuyến giữa hai biến số. Một giá trị ngoại lai có thể làm biến dạng kết quả, dẫn đến một hệ số không phản ánh đúng mối quan hệ thực sự giữa hai biến. Ngược lại, Spearman có khả năng nắm bắt tốt hơn mối quan hệ phi tuyến tính và ít bị ảnh hưởng bởi thứ hạng, cho kết quả về mối quan hệ hoàn hảo (bằng 1) giữa hai biến.

Do đó, hệ số Spearman thường được sử dụng trong các lĩnh vực như khoa học xã hội, y học và tâm lý học, nơi mà mối quan hệ giữa các biến không nhất thiết là tuyến tính và dữ liệu có thể chứa giá trị ngoại lai. Vì vậy, khóa luận lựa chọn Spearman để tìm mối quan hệ giữa tri thức của tác vụ trong từ và độ quan trọng của từ ấy trong dự đoán SRL. Cụ thể, hệ số giúp xác định mức độ tương quan giữa tri thức trong không gian latent (một không gian được xây dựng dựa trên dữ liệu học được từ các chuỗi giá trị ước lượng tri thức về PAS) và độ ảnh hưởng của từ bằng phương pháp Smart Substitution. Điều này đảm bảo kết quả khách quan, giảm thiểu sai sót, và tăng độ chính xác cao trong việc dự đoán SRL dựa trên tri thức PAS.

Bảng 4. *Thuật toán Tính các điểm số tầm quan trọng***.**

Thuật toán. Tính toán điểm số quan trọng

1: **Function: GET_IMPORTANCE_SCORE(D_origin, D_perturb, label_map)**
Đầu vào:
- *D_origin*: DataFrame chứa dữ liệu các câu văn bản gốc gồm các thuộc tính uid, pred, logit, label
- *D_perturb*: DataFrame chứa dữ liệu các câu văn bản được làm nhiễu gồm các thuộc tính uid, pred, logit, label

2: *D* = Dict() // Khởi tạo Dictionary score để lưu điểm tầm quan trọng theo đối số.
3: *L_score* = []; // Khởi tạo Danh sách lưu tất cả điểm tầm quan trọng.
4: // Duyệt qua tất cả các câu trong dữ liệu gốc *D_origin*
5: **for** idx_origin ∈ (0, length(*D_origin*) - 1) **do**

```
        // Duyệt qua tất cả các câu trong dữ liệu xáo trộn D_perturb để lấy ra các câu dữ liệu nhiễu có cùng uid với câu gốc, tức tìm tất cả câu nhiễu từ một câu gốc.
    for idx_perturb ∈ (0, length(D_perturb) - 1) do
        // Nếu uid của câu gốc trùng với uid của dữ liệu xáo trộn sẽ thực hiện việc tính toán mức độ ảnh hưởng và độ hữu ích của từ.
        if D_origin.uid[idx_origin] = D_perturb.uid[idx_perturb] do
            // Lấy ra danh sách các đối số có dự đoán thay đổi so với nhãn.
            changed_args = Get Changed Args(D_perturb.pred, D_perturb.label);
            inf_score, w_inf = Influence Score (D_origin.logit[idx_origin], D_perturb.logit[idx_perturb], index_at_changed_arg);  // Gọi hàm tính Influent score
            // Gọi hàm tính Relevance Score.
            rel_score, w_rel = Relevance Score (D_origin.logit[idx_origin], D_perturb.logit[idx_perturb], D_origin.label[idx_origin], D_origin.pred[idx_origin], D_perturb.pred[idx_perturb]);
            brier_score = 1 – Brier Score Loss (D_origin.label[idx_origin], D_origin.logit[idx_origin], L_map); // Gọi hàm tính Brier score.
            score = {
                "uid": D_origin.uid[idx_origin],
                "influence": sum(inf_score)/ sum(w_inf) if sum(w_inf) ≠ 0 else 0,
                "relevance": sum(rel_score)/ sum(w_rel) if sum(w_rel) ≠ 0 else 0,
                "brier_score": brier_score}
            L_score.append(score) // Lưu score vào L_score
            // Lưu score theo nhóm đối số để phục vụ việc đánh giá.
            for all arg in changed_args do
                if arg not in D then D[arg] = []; end if
                D[arg].append(score);
            end for
        end if
    end for
end for
return D, L_score
end function
```

Chương 5: KẾT QUẢ THỰC NGHIỆM VÀ THẢO LUẬN

5.1 Dữ liệu thực nghiệm

5.1.1 Giới thiệu chung

Bộ dữ liệu được sử dụng xuyên suốt quá trình thực nghiệm của khóa luận là PASBio+ [4]. Bộ dữ liệu được hoàn thiện vào năm 2022 bởi tác giả Tuấn Nguyên Hoài Đức cùng các cộng sự tại khoa Công nghệ Thông tin, trường Đại học Khoa học Tự nhiên – ĐHQG TPHCM, Việt Nam. Bộ dữ liệu được phát triển dựa trên bộ ngữ liệu PASBio [5], vốn được xây dựng chuyên biệt cho tác vụ SRL trên bộ khung đối số PAS dành riêng cho các thuật ngữ chuyên ngành y sinh, phù hợp với tác vụ giải thích của khóa luận.

Bộ dữ liệu PASBio+ được các chuyên gia Y sinh ở Đài Loan đánh giá, kiểm định và được trích xuất từ các mục tóm tắt của các bài báo trên tạp chí trong cơ sở dữ liệu chuyên về lĩnh vực y học MEDLINE. Bộ dữ liệu gồm 35 tệp tương ứng với 35 bộ khung đối số nhưng chỉ với 29 động từ chính, bởi một số động từ trong lĩnh vực Y sinh mang nhiều hơn một bộ khung đối số. Mỗi bộ khung PAS bao gồm hai thành phần chính: vị ngữ và các đối số xoay quanh vị ngữ. ***Bảng 5*** là minh họa bộ khung đối số của động từ "**alter**":

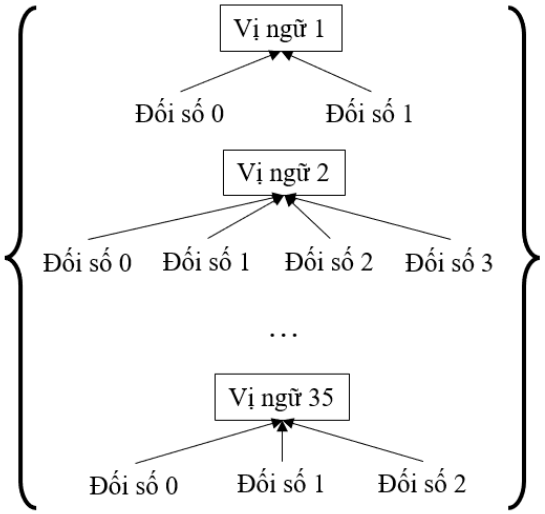
Bảng 5. *Khung đối số của động từ "ALTER".*

Động từ "ALTER"	
Ví dụ: "The codon (CTC) normally associated with Leu-93 in the transcortin polypeptide will be **altered** by the mutation located within exon 2 to a codon (CAC) for histidine in the variant genes."	
Đối số 0: Nguyên nhân gây ra sự thay đổi, có thể do đột biến, một protein…	the mutation located within exon 2
Đối số 1: Đối tượng bị thay đổi, có thể là một codon, exon…	the codon

Đối số 2: Trạng thái sau thay đổi	to a codon (CAC)
Đối số 3: Trạng thái trước thay đổi	the codon (CTC)
Đối số 4: Vị trí xảy ra sự thay đổi, có thể là mô, bào quan…	Không có.

5.1.2 Chi tiết bộ dữ liệu

PASBio+ gồm tất cả 28439 câu với 35 bộ khung đối số nhưng chỉ có 29 động từ chính, vì một số động từ có nhiều hơn một ngữ nghĩa khi được đặt vào ngữ cảnh khác nhau, do đó có nhiều hơn một bộ khung đối số, và vai trò của các đối số sẽ quyết định ngữ nghĩa của động từ ấy. Các đối số sẽ xoay quanh động từ chính và bổ nghĩa cho động từ chính. ***Hình 8*** biểu diễn sơ đồ cấu trúc chung của bộ ngữ liệu.



Hình 8. *Sơ đồ cấu trúc chung của bộ ngữ liệu.*

5.2 Kịch bản thực nghiệm

Với toàn bộ bộ ngữ liệu gồm 28439 câu sẽ được dùng để tinh chỉnh mô hình BioBert trên tác vụ Gán nhãn Ngữ nghĩa được chia với tỉ lệ 60%, 20%, 20% tương ứng cho tập huấn luyện (train), tập kiểm thử (validate) và tập kiểm tra (test). Với điểm số F1 của mô hình sau khi finetune đạt 81%, cho thấy độ đáng tin cậy của mô hình trên tác vụ SRL cần giải thích.

Tuy nhiên, để đơn giản hóa việc tính toán tầm quan trọng của đặc trưng từ lên dự đoán của bài toán. Khóa luận quyết định chỉ sử dụng 6752 câu được lấy mẫu theo phương pháp phân tầng để đại diện chính xác cho cấu trúc tổng thể ban đầu của bộ dữ liệu trong mỗi nhóm động từ. Tập dữ liệu nhỏ này sẽ được dùng để thực nghiệm cho cả phương

pháp Smart Substitution và kiểm tra sự hiện hữu của tri thức PAS trong điểm số tầm quan trọng được tham khảo từ nghiên cứu [85].

Từ bộ dữ liệu 6752 câu cho 35 động từ, khóa luận thực hiện việc phân tách các thực từ trong câu và đánh dấu vị trí các thực từ có vai trò ngữ pháp quan trọng để thực hiện việc thay thế. Chẳng hạn, câu đầu vào gồm 5 thực từ bao gồm cả động từ, danh từ, tính từ, trạng từ thì sẽ có 5 câu được tạo ra bằng cách thay thế tìm từ thay thế mới lần lượt cho 5 thực từ ấy. Tương tự như vậy, khóa luận có được tổng cộng 83232 câu cần được thay thế cho 35 động từ. ***Bảng 6*** thể hiện số liệu chi tiết.

Bảng 6. *Thống kê số câu trước và sau khi nhiễu loạn của mỗi nhóm động từ.*

Động từ	Số câu ban đầu	Số câu được tạo ra để thay thế
Abolish	102	1024
Alter	154	1568
Begin 1	192	2208
Begin 2	247	3520
Block	176	2304
Catalyse	127	2048
Confer	245	3008
Decrease 1	263	3232
Decrease 2	52	544
Delete	83	928
Develop	168	2368
Disrupt	67	800
Eliminate	91	896
Encode	99	1184
Express	150	1632
Generate	99	992
Inhibit	157	1664

Initiate	176	2304
Lead	108	1408
Lose	195	2496
Modify	241	2912
Mutate	480	6336
Proliferate	173	2304
Recognize	78	928
Result	312	3648
Skip	205	2912
Splice 1	279	3296
Splice 2	240	2592
Transcribe	362	3808
Transform 1	140	1024
Transform 2	204	2016
Translate 1	238	3232
Translate 2	273	3008
Translate 3	281	4544
Truncate	295	4544
Tổng	**6752**	**83232**

5.2.1 Mô tả thực nghiệm

Phương pháp đề xuất Smart Substitution: Bộ dữ liệu xáo trộn mới được tạo ra từ phương pháp đề xuất sẽ được dùng để quan sát hành vi cục bộ của mô hình xung quanh từ mới được xáo trộn và đánh giá độ ảnh hưởng và tính hữu tính của từng thực từ đối với từng đối số trong mỗi nhóm động từ của bộ dữ liệu xáo trộn của đặc trưng từ ấy thông qua các đại lượng Influence, Relevance và Competence sẽ được trình bày sau đây. Sau đó, khóa luận tiến hành biểu diễn sự đồng điệu giữa năng lực (competence) của mô hình trong mỗi nhóm với điểm số (1-Brier score loss) tương ứng.

Đồng thời để khẳng định mức độ cải tiến của phương pháp so với các kỹ thuật nhiễu loạn truyền thống như xóa hay che token. Khóa luận cũng tiến hành biểu diễn sự đồng điệu giữa năng lực của mô hình với (1 – Brier score loss) cho mỗi kỹ thuật truyền thống nêu trên để làm baseline so sánh với kỹ thuật Smart Substitution mà khóa luận đề xuất.

Đối với nhiệm vụ kiểm tra có hay không sự tồn tại về tầm quan trọng của từ trong không gian latent, khóa luận tiến hành tìm ra hệ số tương quan Spearman giữa chuỗi độ ảnh hưởng của toàn bộ thực từ với các chuỗi $\hbar_1$, $\hbar_2$ được xây dựng từ hai giả thuyết. Tương ứng với mỗi giả thuyết sẽ được 3 chuỗi $\hbar_1$ và 3 chuỗi $\hbar_2$. Từ đó khóa luận tiến hành so sánh đánh giá với 6 chuỗi $\hbar$ trên. Kết quả được trình bày ở **Mục 5.3**.

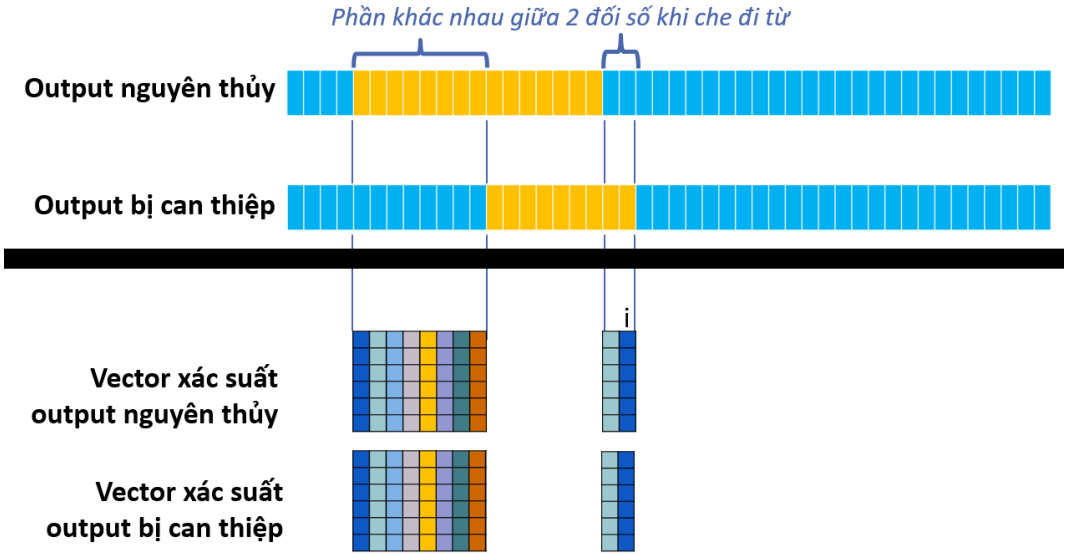
5.2.2 Phương pháp đánh giá

- Đại lượng *Influence* [85]

Để đánh giá độ quan trọng từ trong mỗi dự đoán, nguyên liệu cần sử dụng là vector xác suất của dự đoán trước và sau khi thực hiện Smart Substitution. Nhưng thông thường khi nói đến tầm quan trọng của từ, người ta thường đánh giá mức độ ảnh hưởng (*Influence*) của đặc trưng từ ấy lên dự đoán các đối số của mô hình. Sự góp mặt của từ ấy đã ảnh hưởng như thế nào đến hành vi của mô hình, việc ảnh hưởng này nhằm ủng hộ hay phản đối dự đoán của mô hình. Như vậy, sự hiện diện ấy tác động đến hai hướng của dự đoán, hoặc là dự đoán sẽ bị đảo nhãn, hoặc là dự đoán không đảo nhãn. Dự đoán bị đảo nhãn tức trong dự đoán của đầu vào nguyên thủy và dự đoán của đầu vào can thiệp có sự chuyển đổi từ nhãn IN (tức trong đối số) sang nhãn OUT (nằm ngoài đối số ban đầu) và ngược lại. Dự đoán không đảo nhãn thể hiện rằng dự đoán của đầu vào trước và sau bị can thiệp vẫn nằm trong nhãn IN. Bên cạnh đó, *Influence* dương còn cho biết rằng đặc trưng này có ảnh hưởng đến việc ủng hộ dự đoán của mô hình và *Influence* âm cho thấy đặc trưng ấy đang nhằm mục đích chống đối lại dự đoán của mô hình.

Gọi P là vector xác suất của mô hình cho một mẫu $x \in D$, với D là bộ dữ liệu nguyên thủy. $\tilde{P}$ là vector xác suất của mẫu $\tilde{x} \in \tilde{D}$ trong bộ dữ liệu xáo trộn.

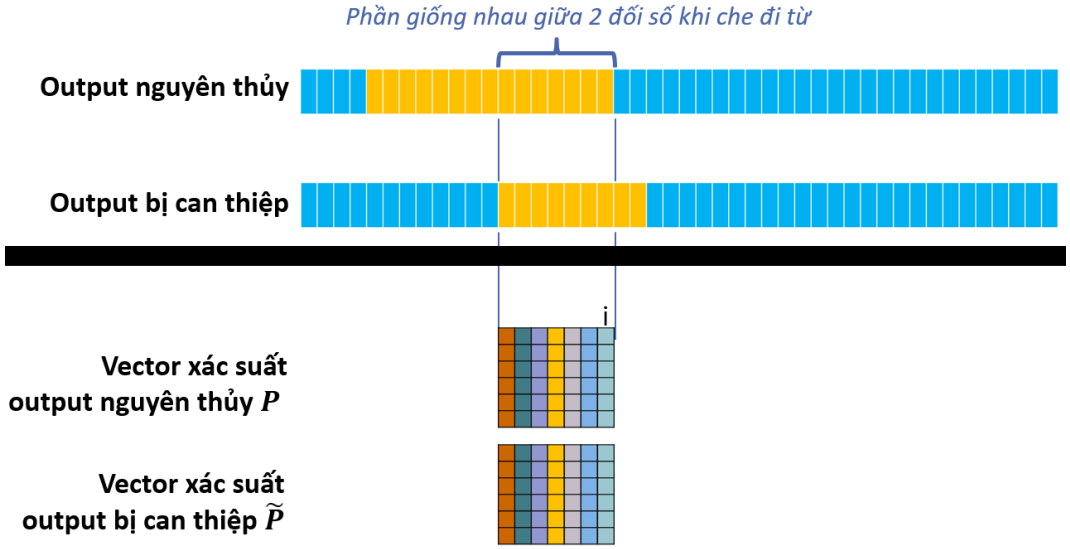
Đối với trường hợp bị đảo nhãn (***Hình 9***), *Inf* của từng token i trong câu được tính như sau:



Hình 9. *Minh họa Trường hợp Dự đoán bị Đảo nhãn.*

$$Inf_i = \left(\frac{P(i.y)-\tilde{P}(i.y)}{Max(P(i.y),\tilde{P}(i.y))} + \frac{\tilde{P}(i.\tilde{y})-P(i.\tilde{y})}{Max(\tilde{P}(i.\tilde{y}),P(i.\tilde{y}))}\right) \quad (21)$$

Đối với trường hợp không đảo nhãn (***Hình 10***):



Hình 10. *Minh họa Trường hợp Không Đảo nhãn.*

$$Inf_i = \left(\frac{P(i.y) - \tilde{P}(i.y)}{Max(P(i.y), \tilde{P}(i.y))}\right) \quad (22)$$

Như vậy, công thức cho mức độ ảnh hưởng của từ w đối với toàn bộ dự đoán p (tức là 1 đối số trong câu) trong SRL model $\mathcal{M}$ sẽ là:

$$_{w}Inf(p) = \frac{\sum_{i \in w} inf_i}{\sum_{i \in w} w_{rel}(i)} \quad (23)$$

$$\text{Trong đó, } w_{rel}(i) = \begin{cases} 2 \; if \; i.y \neq i.\tilde{y} \\ 1 \; if \; i.y = i.\tilde{y} \end{cases}$$

- Đại lượng *Relevance* [85]

Ngoài khía cạnh về mức độ ảnh hưởng của đặc trưng thì còn có mức độ hữu ích (*Relevance*) của đặc trưng từ lên dự đoán. Nếu *Inf* cho biết đặc trưng ấy ảnh hưởng bao nhiêu phần trăm vào dự đoán của mô hình thì *Rel* chỉ ra đặc trưng ấy làm cho mô hình chính xác bao nhiêu phần trăm. Như vậy, *Rel* sẽ giúp đo lường những từ mà giúp dự đoán mô hình gần giống với nhãn đối số "vàng" được bao nhiêu phần trăm. Chẳng hạn lúc có sự hiện diện của từ ấy thì mô hình dự đoán được 90%, nhưng khi từ ấy bị xáo trộn thành từ khác thì dự đoán mô hình chỉ còn 50%, như vậy 40% chính là *độ hữu ích* của từ ấy đối với dự đoán đang xét, chính từ ấy đã giúp mô hình từ 50% lên 90%. Độ hữu ích của mỗi token t trong Jugdment Space (JudSp) (***Hình 11***) được tính toán bằng sự thay đổi tăng giảm gây ra bởi sự hiện diện của từ w lên các vector xác suất:

$$_{w}Rel(t) = \begin{cases} \frac{^{w}P^{\uparrow}(t.\hat{y})}{P^{Max}(t.\hat{y})} & if \; t.y = t.\tilde{y} = t.\hat{y} \\ \frac{1}{2}\left(\frac{^{w}P^{\uparrow}(t.\hat{y})}{P^{Max}(t.\hat{y})} + \frac{^{w}P^{\downarrow}(t.\tilde{y})}{P^{Max}(t.\tilde{y})}\right) & if \; t.\tilde{y} \neq t.y = t.\hat{y} \\ \frac{1}{2}\left(\frac{^{w}P^{\uparrow}(t.\hat{y})}{P^{Max}(t.\hat{y})} + \frac{^{w}P^{\downarrow}(t.y)}{P^{Max}(t.y)}\right) & if \; t.y \neq t.\tilde{y} = t.\hat{y} \\ \frac{1}{3}\left(\frac{^{w}P^{\uparrow}(t.\hat{y})}{P^{Max}(t.\hat{y})} + \frac{^{w}P^{\downarrow}(t.y)}{P^{Max}(t.y)} + \frac{^{w}P^{\downarrow}(t.\tilde{y})}{P^{Max}(t.\tilde{y})}\right) & if \begin{cases} t.\tilde{y} \neq t.\hat{y} \\ t.y \neq t.\hat{y} \end{cases} \end{cases} \quad (24)$$

Trong đó: $^{w}P^{\uparrow}(t.\hat{y}) = P(t.\hat{y}) - \tilde{P}(t.\hat{y})$

$^{w}P^{\downarrow}(t.y) = \tilde{P}(t.y) - P(t.y)$

$$ {}^{w}P^{\downarrow}(t.\tilde{y}) = \tilde{P}(t.\tilde{y}) - P(t.\tilde{y}) $$

Với $t.y$ là nhãn gold của token t

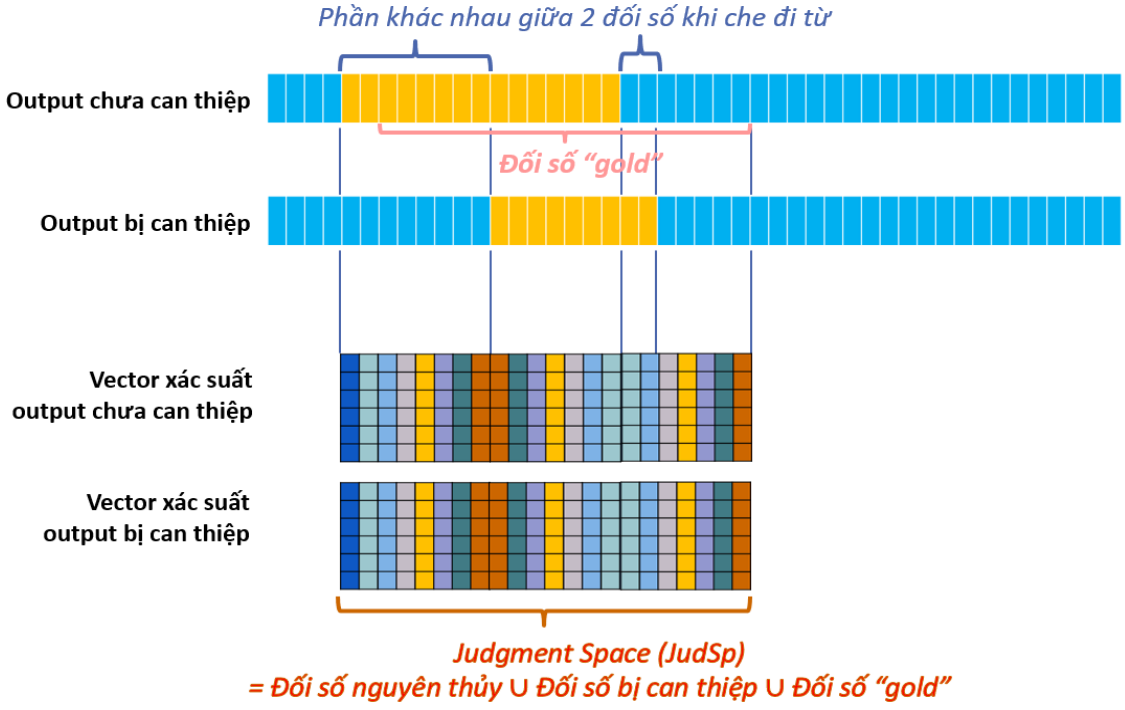
$t.\hat{y}$ là nhãn được dự đoán của token t khi chưa được can thiệp

$t.\tilde{y}$ là nhãn được dự đoán của token t khi được can thiệp

Tương tự, $P(t.y)$ là xác suất của dự đoán lúc chưa can thiệp

$\tilde{P}(t.\hat{y})$ là xác suất của nhãn gold trong dự đoán lúc đã can thiệp

$\tilde{P}(t.y)$ là xác suất của nhãn nguyên thủy trong kết quả đã can thiệp



Hình 11. *Minh họa về Judgment Space (JudSp).*

Độ hữu ích của từ đối với toàn bộ dự đoán $\mathcal{p}$ (tức là 1 argument trong câu) trong SRL model $\mathcal{M}$ sẽ là:

$$ {}_{w}Rel(\mathcal{p}) = \frac{\sum_{t \in {}_{w}JudSp(\mathcal{p})} w_{rel}(t) \cdot {}_{w}Rel(t)}{\sum_{t \in {}_{w}JudSp(\mathcal{p})} w_{rel}(t)} \quad (25) $$

Trong đó, trọng số của từng token t, $w_{rel}(t)$ được tính toán như sau:

$$ w_{rel}(t) = \begin{cases} 1 & if \ \ (t.y = t.\tilde{y} = t.\hat{y}) \vee (t.y \neq t.\hat{y} \wedge t.\tilde{y} \neq t.\hat{y}) \\ 2 & if \ otherwise \end{cases} $$

- Đại lượng *Competence(Comp)* [85]

Ngoài việc khai thác sức ảnh hưởng và mức độ hữu ích của đặc trưng lên dự đoán thì khóa luận còn sử dụng thêm đại lượng *Competence*. *Competence* chính là năng lực của mô hình SRL, năng lực sử dụng đặc trưng. *Comp* cao chứng tỏ mô hình biết cách sử dụng những đặc trưng hữu ích giúp cải thiện độ chính xác của mô hình và hạn chế sử dụng những đặc trưng có hại, làm suy giảm độ chính xác của mô hình. Những đặc trưng có độ hữu ích cao thì sẽ được mô hình sử dụng nhiều, nghĩa là để nó ảnh hưởng lên dự đoán nhiều (*Inf* cao). Nếu có độ hữu ích thấp thì làm cho dự đoán xa rời với dự đoán chuẩn "vàng", nó không hữu ích thì cho mức độ ảnh hưởng thấp (*Inf* thấp). Do đó, nếu mô hình có năng lực (*Comp*) cao thì nó sẽ đủ thông minh để hạn chế sử dụng những đặc trưng cho tầm ảnh hưởng thấp. Vì thế, năng lực của mô hình sẽ được đánh giá bằng hệ số tương quan Spearman giữa *Inf* và *Rel*. Hệ số tương quan thể hiện sự đồng điệu giữa hai biến số *Inf* và *Rel*. Nếu *Rel* cao thì *Inf* cũng cao và ngược lại.

Năng lực của mô hình $\mathcal{M}$ khi sử dụng mức độ quan trọng của từ w để đưa ra tập dự đoán $\mathcal{P}$={ p_1, p_2, …, p_n }, được ước lượng bởi phương pháp giải thích $\mathcal{E}$ là hệ số tương quan Spearman (Spearman correlation coefficient - r_s) giữa *Influence* và *Relevance* khi biến thiên trên tập dữ liệu.

$$_{\mathcal{M}}^{\mathcal{E}}Comp(\mathcal{P}, w) = r_s(\{|\mathcal{E}.Inf(\mathcal{M}, p_i, w)|,\ \mathcal{E}.Rel(\mathcal{M}, p_i, w)\}_{i=1}^{n}) \quad (26)$$

- Độ đo *Brier score loss*

Để đo lường tầm quan trọng của đặc trưng từ tác động lên dự đoán của mô hình, khóa luận tiến hành tính toán phần trăm năng lực dự đoán của mô hình, thay vì dùng điểm số F1 để đo đạt hiệu suất của mô hình, khóa luận sẽ sử dụng điểm số hàm mất mát Brier với các ưu điểm sau. Thứ nhất, Brier score đo lường chính xác của các dự đoán xác suất, cung cấp thông tin chi tiết về mức độ tin cậy của mô hình cho từng dự đoán cụ thể thay vì chỉ dựa trên phân loại nhị phân như điểm F1 đã làm. Đối với đề tài của khóa luận cần cung cấp chính xác về tác động của đặc trưng lên hiệu suất của mô hình, khóa luận muốn

đánh giá dựa trên các dự đoán xác suất thay vì chỉ dựa trên phân loại nhãn. Chẳng hạn, xét ví dụ minh họa sau để thấy sự khác biệt rõ rệt về điểm số Brier và điểm số F1:

$$Brier\ score\ loss = \frac{1}{N}.\sum_{i=1}^{N}(f_i - o_i)^2 \quad (27)$$

Trong đó, f_i là xác suất dự đoán, và o_i là giá trị thực tế, N là tổng số mẫu.

5.3 Kết quả thực nghiệm

5.3.1 Phương pháp Smart Substitution

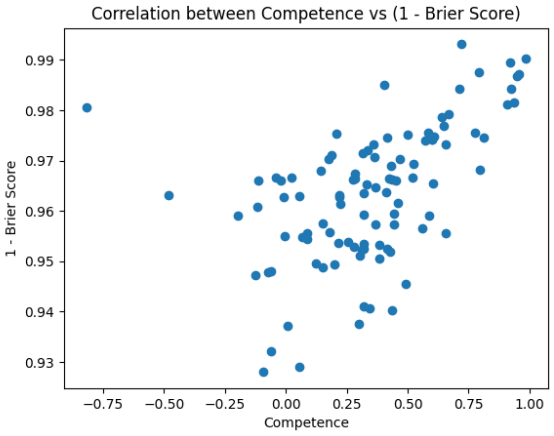
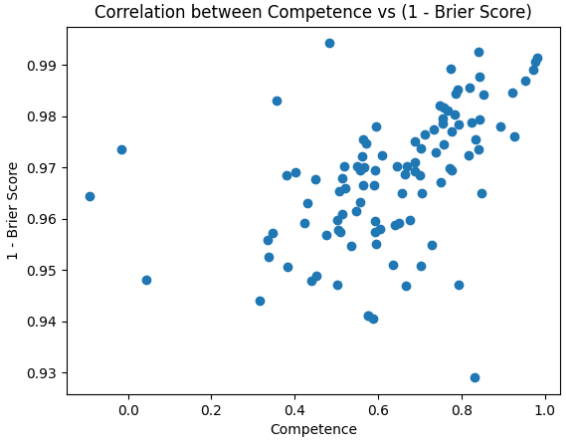
Thực hiện 8 phương pháp thay thế và kết quả hệ số tương quan giữa Competence và (1 – Brier score loss) của các phương pháp được thể hiện trong **Bảng 7**.

Bảng 7. *Bảng số liệu thống kê kết quả 8 phương pháp.*

		Độ tương đồng gần 0 (pos)	**Độ tương đồng gần -1 (neg)**
SUM	**cosine-module**	0.78	0.72
	cosine	0.78	0.71
AVG	**cosine-module**	0.78	0.73
	cosine	0.78	0.72

- Kết quả đối với phương pháp làm nhiễu bằng cách thay bằng token [MASK] hay xóa token được thể hiện trong ***Bảng 8***.

Bảng 8. *Kết quả của phương pháp mask và xóa token.*

Phương pháp thay bằng token [MASK]	**Phương pháp xóa token**
 Tương quan Spearman: 0.57	 Tương quan Spearman: 0.58

Bảng 7 thể hiện độ tương quan giữa năng lực của mô hình và điểm (1- Brier score) của 8 phương pháp thay thế từ thông minh cho thấy:

Đối với phương pháp chọn từ thay thế xa nghĩa (pos) và trái nghĩa (neg) và cho thấy tất cả các phương pháp chọn từ có liên quan đến từ xa nghĩa bất kể vector biểu diễn được lấy tổng hay trung bình và bất kể sử dụng độ tương đồng nào (cosine, cosine module) thì cũng đều có mối tương quan cao ấn tượng (0.78) so với các phương pháp liên quan đến chọn từ trái nghĩa (cao nhất chỉ 0.73).

Điều này cho thấy khi thay thế từ không liên quan về ngữ nghĩa với từ gốc sẽ cho lời giải thích trung thực hơn so với khi thay bằng từ ngược nghĩa với từ gốc. Điều đó thể hiện ở chỗ lời giải thích này cho phép ước lượng năng lực của mô hình có tương quan cao hơn với hiệu quả dự đoán thực tế của mô hình.

Đối với phương pháp đo tương đồng giữa từ gốc với từ ứng viên sử dụng độ đo cosine hay cosine module thì giá trị cosine module có xu hướng cao hơn hoặc tương đương với

giá trị cosine, đặc biệt là với các từ trái nghĩa. Điều này cho thấy cosine module có tính ổn định hơn khi có cân nhắc đến độ dài vector. Cụ thể cùng là một phương thức lấy trung bình (AVG) vector biểu diễn và từ ứng viên đều trái nghĩa (neg) với từ gốc thì avg_neg_cos_module (0.73) vượt trội hơn avg_neg_cos (0.71). Tương tự cho phương thức lấy tổng (SUM) vector biểu diễn và từ ứng viên đều trái nghĩa (neg) với từ gốc thì sum_neg_cos_module (0.72) cũng cao hơn sum_neg_cos (0.71). Còn đối với từ không liên quan đến nghĩa, cho kết quả tương đối giống nhau giữa cosine và cosine module (đều là 0.78).

Điều này cho thấy việc sử dụng cosine module để thay thế các từ trái nghĩa có thể giúp giải thích tầm quan trọng của đặc trưng một cách trung thực hơn, nhưng không hữu ích hơn một cách rõ rệt so với cosine khi sử dụng để thay thế các từ xa nghĩa.

Đối với việc lấy trung bình hay tổng các vector biểu diễn token để làm vector biểu diễn từ thì kết quả thu được không có sự khác biệt lớn. Cụ thể, trong cùng một phương pháp chọn từ trái nghĩa và sử dụng độ đo cosine module thì kết quả thu được cho hai cách lấy trung bình và tổng lần lượt là avg_neg_cos_module (0.73) và sum_neg_cos_module (0.72). Tương tự cho việc chọn từ trái nghĩa và sử dụng độ đo cosine thì kết quả tương đương nhau avg_neg_cos (0.71) và sum_neg_cos (0.71).

Kết luận:

Các phương pháp chọn từ xa nghĩa (cosine gần 0, pos) cho kết quả tốt hơn so với các phương pháp chọn từ trái nghĩa (cosine gần -1, neg), điều đó giúp cho phương pháp Smart Substitution có khả năng giải thích được tốt hơn hành vi của mô hình khi xáo trộn bằng từ xa nghĩa. Việc sử dụng độ tương đồng cosine module thay cho cosine có thể cải thiện hiệu quả giải thích khi thay thế bằng các các từ trái nghĩa, nhưng ít ảnh hưởng đến các từ xa nghĩa. Tuy nhiên, không có sự khác biệt lớn giữa các phương pháp tính tổng và trung bình các vector biểu diễn token làm vector biểu diễn từ cần giải thích. Bên cạnh đó, khi so sánh với phương pháp xáo trộn dữ liệu bằng cách xóa hay thay bằng token

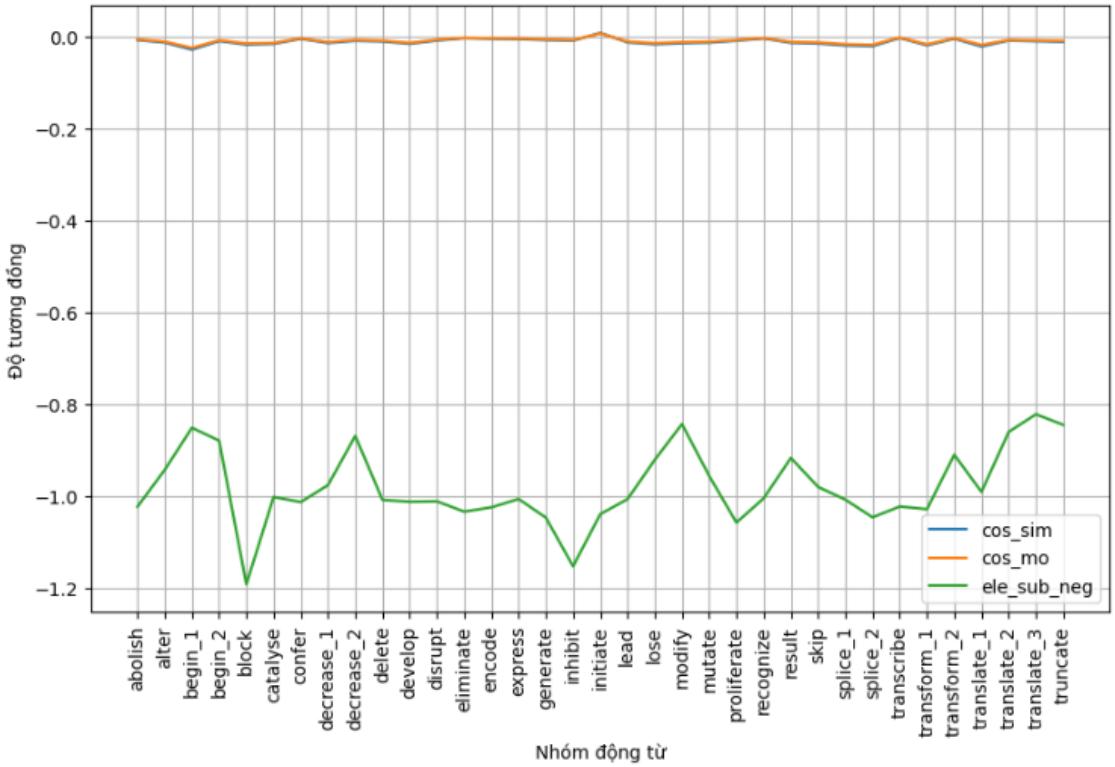
[MASK], tất cả phương thức thay thế của phương pháp Smart Substitution đều cho kết quả vượt trội hơn hẳn. Kết quả thấp nhất của phương pháp Smart Substitution sum_neg_cos là 0.71 cao hơn hẳn 0.57 hay 0.58. (**Bảng 7**). Như vậy, phương pháp Smart Substitution thực hiện việc nhiễu loạn dữ liệu là phương pháp hiệu quả hơn rõ rệt so với cái phương pháp nhiễu loạn thông thường, nhằm định lượng được giá trị *influence* và *relevance* trung thực nhất có thể cho mỗi thực từ đang xét.

5.3.2 Sự hiện hữu tầm quan trọng trong không gian vector

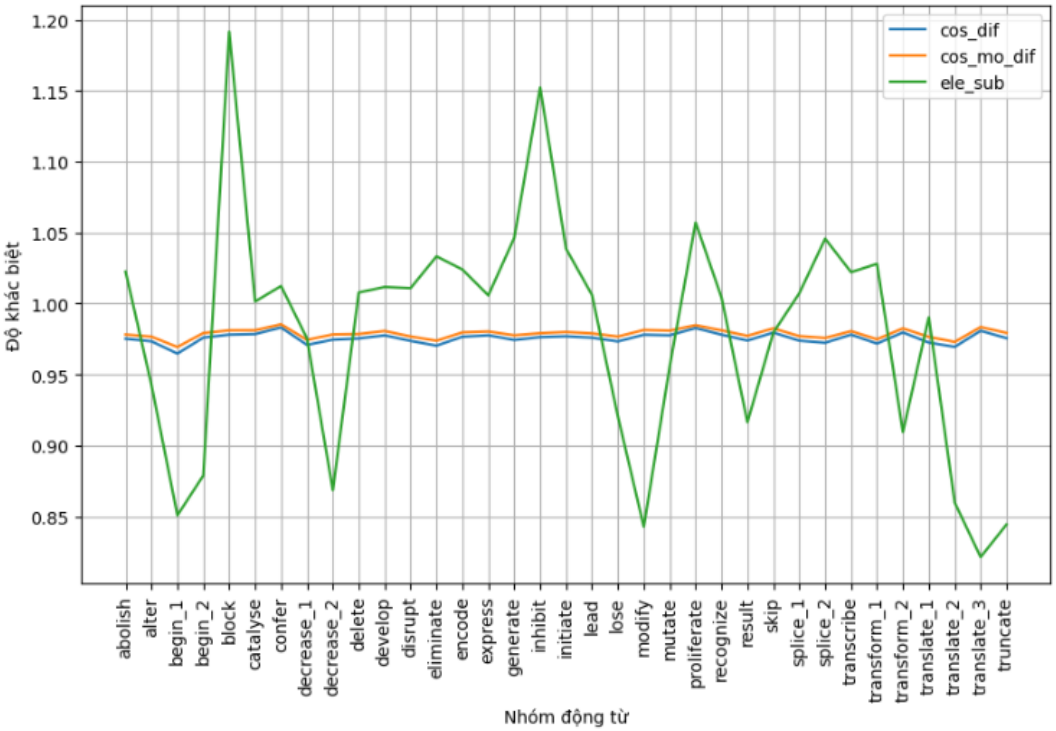
Để ước lượng tri thức về PAS được mã hóa trong vector biểu diễn từ trước và sau khi tinh chỉnh, khóa luận thực tính độ tương quan (được nêu trong Vấn đề mở 2) giữa hai vector này.

Đối với Giả thuyết 1 được định lượng bởi độ tương đồng cosine, cosine module và -|element-wise substraction|, kết quả thu được trong ***Hình 12.***

Đối với Giả thuyết 2, kết quả thu được cho độ khác biệt (1 – cosine), (1 – cosine_module) và |(element-wise substraction)| trong ***Hình 13***.



Hình 12. *Kết quả ước lượng tri thức PAS theo Giả thuyết 1.*



Hình 13. *Kết quả ước lượng tri thức PAS theo Giả thuyết 2.*

Hình 12 Từ kết quả thực nghiệm cho thấy giá trị về độ tương đồng ở Giả thuyết 1 giữa hai vector biểu diễn trước và sau khi finetune của mô hình rất thấp, tất cả các nhóm động từ chỉ giao động quanh 0 đối với độ đo cosine và cosine module. Thậm chí với độ đo -|element-wise substraction| giá trị độ tương đồng nằm trong khoảng -1.2 đến -0.8.

Đối lập hoàn toàn với độ tương đồng ở Giả thuyết 1, Giả thuyết 2 chứng minh được vector biểu diễn trước và sau khi finetune có khác biệt vượt trội (giá trị độ khác biệt giao động gần 0.95). Thậm chí đối với độ đo |(element-wise substraction)|, giá trị độ chênh lệch rất cao (từ 0.85 đến gần 1.2). Điều đó chứng tỏ rằng quá trình finetune đã dạy cho mô hình tập trung tích lũy rất nhiều tri thức về PAS trong vector biểu diễn. (***Hình 13***)

Kết quả thực nghiệm về sự hiện hữu tầm quan trọng của từ trong vector biểu diễn của từ trong ***Bảng 9***.

Bảng 9. *Mối tương quan giữa các chuỗi tri thức PAS và điểm Inf, Rel.*

	cos	**cosmo**	**-|ele-wise sub|**	**1-cos**	**1-cosmo**	**|ele-wise sub|**
Influence	-0.016	-0.023	-0.266	-0.06	-0.079	0.27
Relevance	0.033	0.028	-0.256	-0.003	-0.021	0.26

Bảng 9. Kết quả cho quá trình thực nghiệm về kiểm tra liệu điểm số tầm quan trọng đặc trưng được khóa luận sử dụng có phản ánh được tri thức về PAS trong vector biểu diễn. Cụ thể, mối tương quan giữa giá trị *Inf, Rel* cho cả 6 chuỗi tri thức PAS được xây dựng từ Giả thuyết 1 và Giả thuyết 2 đều rất thấp, chỉ giao động quanh quẩn từ -0.01 đến 0.3. Như vậy, ngay cả độ tương quan dương cao nhất (0.27) cũng không cao hơn ngưỡng 0.5 cho thấy vector biểu diễn của một thực từ không mã hóa gì về tầm quan trọng (*Influence, Relevance*) của từ đó đối với tác vụ SRL đang xét.

Chương 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1 Kết luận

6.1.1 Cở sở lý thuyết đã tìm hiểu

- Tác vụ Gán nhãn Ngữ nghĩa trong Y sinh.
- Nghiên cứu về bài toán XAI và các tầm quan trọng của tính Khả diễn giải trong các mô hình học sâu.
- Các kỹ thuật giải thích dựa trên tầm quan trọng của đặc trưng trong bài toán XAI.

6.1.2 Đóng góp của khóa luận

- Khóa luận đã đề xuất một phương pháp mới trong việc làm nhiễu dữ liệu, là công đoạn rất quan trọng của hướng tiếp cận giải thích hậu nghiệm dựa trên tầm quan trọng của đặc trưng. Phương pháp làm nhiễu dữ liệu của khóa luận giúp khắc phục các hạn chế về tính trung thực của lời giải thích mà các phương pháp làm nhiễu dữ liệu phổ biến hiện nay đang gặp phải.
- Kết quả thực nghiệm cho tác vụ SRL trên văn bản Y Sinh cho thấy các phương pháp được đề xuất trong khóa luận này cho ra lời giải thích phù hợp với hiệu quả dự đoán thực tế của mô hình và đạt được độ trung thực vượt trội hơn các phương pháp làm nhiễu dữ liệu hiện có. Bên cạnh đó, khóa luận cũng thực hiện thăm dò kiểm tra sự hiện hữu của tầm quan trọng đặc trưng trong không gian latent và rút ra kết luận từ thực nghiệm.

6.2 Hướng phát triển

- Hiện tại, phương pháp làm nhiễu dữ liệu của khóa luận chỉ xem xét từ loại và ngữ nghĩa của từ khi chọn ứng viên thay thế. Hướng phát triển tương lai sẽ quan tâm thêm các khía cạnh chuyên sâu hơn về ngữ pháp như sự tương hợp giữa chủ từ và vị từ, thì (tense) và cách (voice) của động từ…

- Ngoài ra, việc ứng dụng giá trị Shapley trong giải thích tầm quan trọng của đặc trưng cũng là một hướng nghiên cứu đáng quan tâm. Trong tương lai, khóa luận sẽ được tiếp nối bằng việc nghiên cứu để tích hợp ước lượng Shapley vào các thuật toán đã được đề xuất trong khóa luận này.

Danh mục tài liệu tham khảo

[1] Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. In *Proceedings of the International Conference on Learning Representations*, August 15, 2017. 1–12. . Retrieved October 18, 2023 from https://arxiv.org/abs/1608.04207

[2] Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno-Schneider, and Georg Rehm. 2021. Fine-grained Classification of Political Bias in German News: A Data Set and Initial Experiments. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, August 2021. Association for Computational Linguistics, Stroudsburg, PA, USA, 121–131. . https://doi.org/10.18653/v1/2021.woah-1.13

[3] David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, September 2017. Association for Computational Linguistics, Stroudsburg, PA, USA, 412–421. . https://doi.org/10.18653/v1/D17-1042

[4] Malika Aubakirova and Mohit Bansal. 2016. Interpreting Neural Networks to Improve Politeness Comprehension. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, November 2016. Association for Computational Linguistics, Stroudsburg, PA, USA, 2035–2041. . https://doi.org/10.18653/v1/D16-1216

[5] Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen Fraser. 2022. Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, 2022. Association for Computational Linguistics, Stroudsburg, PA, USA, 80–92. . https://doi.org/10.18653/v1/2022.trustnlp-1.8

[6] Esma Balkir, Isar Nejadgholi, Kathleen Fraser, and Svetlana Kiritchenko. 2022. Necessity and Sufficiency for Explaining Text Classifiers: A Case Study in Hate Speech Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, July 2022. Association for Computational Linguistics, Stroudsburg, PA, USA, 2672–2686. . https://doi.org/10.18653/v1/2022.naacl-main.192

[7] Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and Analysis in Neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 2020. Association for Computational Linguistics, Stroudsburg, PA, USA, 1–5. .

https://doi.org/10.18653/v1/2020.acl-tutorials.1

[8] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March 03, 2021. ACM, New York, NY, USA, 610–623. . https://doi.org/10.1145/3442188.3445922

[9] Milan Bhan, Nina Achache, Victor Legrand, Annabelle Blangero, and Nicolas Chesneau. 2023. Evaluating self-attention interpretability through human-grounded experimental protocol. *arXiv:2303.15190* (March 2023). Retrieved October 18, 2023 from https://arxiv.org/abs/2303.15190

[10] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, January 27, 2020. ACM, New York, NY, USA, 648–657. . https://doi.org/10.1145/3351095.3375624

[11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Proceedings of the Advances in Neural Information Processing Systems*, May 28, 2020. 1877–1901. . Retrieved October 18, 2023 from https://dl.acm.org/doi/abs/10.5555/3495724.3495883

[12] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive Adversarial Networks: High-Recall Explanations for Identifying Personal Attacks in Social Media Posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, November 2018. Association for Computational Linguistics, Stroudsburg, PA, USA, 3497–3507. . https://doi.org/10.18653/v1/D18-1386

[13] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics (Basel)* 8, 8 (July 2019), 832. https://doi.org/10.3390/electronics8080832

[14] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural*

Networks for NLP, August 2019. Association for Computational Linguistics, Stroudsburg, PA, USA, 276–286. . https://doi.org/10.18653/v1/W19-4828

[15] Louis Clouatre, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar. 2022. Local Structure Matters Most: Perturbation Study in NLU. In *Findings of the Association for Computational Linguistics: ACL 2022*, May 2022. Association for Computational Linguistics, Stroudsburg, PA, USA, 3712–3731. . https://doi.org/10.18653/v1/2022.findings-acl.293

[16] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2018. Association for Computational Linguistics, Stroudsburg, PA, USA, 2126–2136. . https://doi.org/10.18653/v1/P18-1198

[17] Danilo Croce, Daniele Rossini, and Roberto Basili. 2019. Auditing Deep Learning processes through Kernel-based Explanatory Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, November 2019. Association for Computational Linguistics, Stroudsburg, PA, USA, 4035–4044. . https://doi.org/10.18653/v1/D19-1415

[18] Evan Crothers, Herna Viktor, and Nathalie Japkowicz. 2023. Faithful to Whom? Questioning Interpretability Measures in NLP. *arXiv:2308.06795* (August 2023). Retrieved October 18, 2023 from https://arxiv.org/abs/2308.06795

[19] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020. Association for Computational Linguistics, 447–459. . Retrieved from https://xainlp2020.github.io/xainlp/

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, June 02, 2019. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. .

[21] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv: Machine Learning* (February 2017). Retrieved October 18, 2023 from https://arxiv.org/abs/1702.08608

[22] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Christopher Bavitz, Samuel J.

Gershman, David O'Brien, Stuart Shieber, Jim Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI Under the Law: The Role of Explanation. *SSRN Electronic Journal* (2017). https://doi.org/10.2139/ssrn.3064761

[23] Kevin Du, Lucas Torroba Hennigen, Niklas Stoehr, Alex Warstadt, and Ryan Cotterell. 2023. Generalizing Backpropagation for Gradient-Based Interpretability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2023. Association for Computational Linguistics, Stroudsburg, PA, USA, 11979–11995. . https://doi.org/10.18653/v1/2023.acl-long.669

[24] Lilian Edwards and Michael Veale. 2017. Slave to the Algorithm? Why a Right to Explanationn is Probably Not the Remedy You are Looking for. *SSRN Electronic Journal* (2017). https://doi.org/10.2139/ssrn.2972855

[25] Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. CausaLM: Causal Model Explanation Through Counterfactual Language Models. *Computational Linguistics* 47, 2 (May 2021), 1–54. https://doi.org/10.1162/coli_a_00404

[26] Nicolas Garneau, Jean-Samuel Leboeuf, and Luc Lamontagne. 2018. Predicting and interpreting embeddings for out of vocabulary words in downstream tasks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, November 2018. Association for Computational Linguistics, Stroudsburg, PA, USA, 331–333. . https://doi.org/10.18653/v1/W18-5439

[27] Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López. 2021. A Survey on Bias in Deep NLP. *Applied Sciences* 11, 7 (April 2021), 3184. https://doi.org/10.3390/app11073184

[28] Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. 2018. Interpreting Recurrent and Attention-Based Neural Models: a Case Study on Natural Language Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, November 2018. Association for Computational Linguistics, Stroudsburg, PA, USA, 4952–4957. . https://doi.org/10.18653/v1/D18-1537

[29] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, October 2018. IEEE, 80–89. . https://doi.org/10.1109/DSAA.2018.00018

[30] Pankaj Gupta and Hinrich Schütze. 2018. LISA: Explaining Recurrent Neural Network Judgments via Layer-wIse Semantic Accumulation and Example to Pattern

Transformation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, November 2018. Association for Computational Linguistics, Stroudsburg, PA, USA, 154–164. . https://doi.org/10.18653/v1/W18-5418

[31] Shiou Tian Hsu, Changsung Moon, Paul Jones, and Nagiza Samatova. 2018. An Interpretable Generative Adversarial Approach to Classification of Latent Entity Relations in Unstructured Sentences. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, April 27, 2018. 5181–5188. . https://doi.org/10.1609/aaai.v32i1.11972

[32] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. Association for Computational Linguistics, Stroudsburg, PA, USA, 4198–4205. . https://doi.org/10.18653/v1/2020.acl-main.386

[33] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North*, 2019. Association for Computational Linguistics, Stroudsburg, PA, USA, 3543–3556. . https://doi.org/10.18653/v1/N19-1357

[34] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020. Curran Associates Inc., 12388–12401. . Retrieved October 18, 2023 from https://dl.acm.org/doi/pdf/10.5555/3495724.3496763

[35] Yichen Jiang, Nitish Joshi, Yen-Chun Chen, and Mohit Bansal. 2019. Explore, Propose, and Assemble: An Interpretable Model for Multi-Hop Reading Comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July 2019. Association for Computational Linguistics, Stroudsburg, PA, USA, 2714–2725. . https://doi.org/10.18653/v1/P19-1261

[36] Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2020. Towards Hierarchical Importance Attribution: Explaining Compositional Semantics for Neural Sequence Models. In *Proceedings of the International Conference on Learning Representations*, November 07, 2020. . Retrieved October 18, 2023 from https://arxiv.org/abs/1911.06194

[37] Judea Pearl. 2001. Direct and Indirect Effects. In *Proceedings of the 17th*

Conference on Uncertainty in Artificial Intelligence, 2001. Morgan Kaufmann Publishers Inc., San Francisco, 411–420. . Retrieved October 18, 2023 from https://ftp.cs.ucla.edu/pub/stat_ser/R273-U

[38] Dongyeop Kang, Varun Gangal, Ang Lu, Zheng Chen, and Eduard Hovy. 2017. Detecting and Explaining Causes From Text For a Time Series Event. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, September 2017. Association for Computational Linguistics, Stroudsburg, PA, USA, 2758–2767. . https://doi.org/10.18653/v1/D17-1292

[39] Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018. Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, June 2018. Association for Computational Linguistics, Stroudsburg, PA, USA, 701–707. . https://doi.org/10.18653/v1/N18-2110

[40] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing Hate Speech Classifiers with Post-hoc Explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020. Association for Computational Linguistics, Stroudsburg, PA, USA, 5435–5442. . https://doi.org/10.18653/v1/2020.acl-main.483

[41] Dongfang Li, Baotian Hu, Qingcai Chen, and Shan He. 2023. Towards Faithful Explanations for Text Classification with Robustness Improvement and Explanation Guided Training. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, July 2023. Association for Computational Linguistics, Stroudsburg, PA, USA, 1–14. . https://doi.org/10.18653/v1/2023.trustnlp-1.1

[42] Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun ACM* 61, 10 (September 2018), 36–43. https://doi.org/10.1145/3233231

[43] Junyu Lu, Chenbin Zhang, Zeying Xie, Guang Ling, Tom Chao Zhou, and Zenglin Xu. 2019. Constructing Interpretive Spatio-Temporal Features for Multi-Turn Responses Selection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July 2019. Association for Computational Linguistics, Stroudsburg, PA, USA, 44–50. . https://doi.org/10.18653/v1/P19-1006

[44] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, December 22, 2017. Curran Associates Inc., 4768–4777. . Retrieved October 18, 2023 from

https://dl.acm.org/doi/10.5555/3295222.3295230

[45] Ling Luo, Xiang Ao, Feiyang Pan, Jin Wang, Tong Zhao, Ningzi Yu, and Qing He. 2018. Beyond Polarity: Interpretable Financial Sentiment Analysis with Hierarchical Query-driven Attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, July 2018. International Joint Conferences on Artificial Intelligence Organization, California, 4244–4250. . https://doi.org/10.24963/ijcai.2018/590

[46] Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. 2022. Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, October 15, 2022. Association for Computational Linguistics, 1731–1751. . Retrieved October 18, 2023 from https://arxiv.org/abs/2110.08412

[47] Andreas Madsen, Siva Reddy, and Sarath Chandar. 2023. Post-hoc Interpretability for Neural NLP: A Survey. *ACM Comput Surv* 55, 8 (August 2023), 1–42. https://doi.org/10.1145/3546577

[48] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. *ACM Comput Surv* 54, 6 (July 2022), 1–35. https://doi.org/10.1145/3457607

[49] Mihály Héder. 2023. Explainable AI: A Brief History of the Concept. In *ERCIM News* . ERCIM EEIG, 9–10. Retrieved October 18, 2023 from https://ercim-news.ercim.eu/en134/special/explainable-ai-a-brief-history-of-the-concept

[50] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif Intell* 267, (February 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[51] Edoardo Mosca, Maximilian Wich, and Georg Groh. 2021. Understanding and Interpreting the Impact of User Context in Hate Speech Detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, June 2021. Association for Computational Linguistics, Stroudsburg, PA, USA, 91–102. . https://doi.org/10.18653/v1/2021.socialnlp-1.8

[52] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, June 2018. Association for Computational Linguistics, Stroudsburg, PA, USA, 1101–1111. . https://doi.org/10.18653/v1/N18-1100

[53] W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs. *arXiv:1801.05453* (January 2018). Retrieved October 18, 2023 from https://arxiv.org/abs/1801.05453

[54] Isar Nejadgholi, Kathleen Fraser, and Svetlana Kiritchenko. 2022. Improving Generalizability in Implicitly Abusive Language Detection with Concept Activation Vectors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, May 2022. Association for Computational Linguistics, Stroudsburg, PA, USA, 5517–5529. . https://doi.org/10.18653/v1/2022.acl-long.378

[55] Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating Robustness and Interpretability of Link Prediction via Adversarial Modifications. In *Proceedings of NAACL-HLT 2019*, June 01, 2019. Association for Computational Linguistics, 3336–3347. . Retrieved October 18, 2023 from http://dx.doi.org/10.18653/v1/N19-1337

[56] Nina Poerner, Hinrich Schütze, and Benjamin Roth. 2018. Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2018. Association for Computational Linguistics, Stroudsburg, PA, USA, 340–350. . https://doi.org/10.18653/v1/P18-1032

[57] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July 2019. Association for Computational Linguistics, Stroudsburg, PA, USA, 4932–4942. . https://doi.org/10.18653/v1/P19-1487

[58] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. Association for Computational Linguistics, Stroudsburg, PA, USA, 3980–3990. . https://doi.org/10.18653/v1/D19-1410

[59] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, June 2016. Association for Computational Linguistics, Stroudsburg, PA, USA, 97–101. . https://doi.org/10.18653/v1/N16-3020

[60] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in

BERTology: What We Know About How BERT Works. *Trans Assoc Comput Linguist* 8, (December 2020), 842–866. https://doi.org/10.1162/tacl_a_00349

[61] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, August 10, 2017. 2662–2670. . Retrieved October 18, 2023 from https://www.ijcai.org/proceedings/2017/371

[62] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1, 5 (May 2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[63] Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July 2019. Association for Computational Linguistics, Stroudsburg, PA, USA, 2931–2951. . https://doi.org/10.18653/v1/P19-1282

[64] Chandan Singh, W. James Murdoch, and Bin Yu. 2018. Hierarchical interpretations for neural network predictions. *ICLR 2019* (June 2018). Retrieved October 18, 2023 from https://arxiv.org/abs/1806.05337

[65] Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, August 2021. Association for Computational Linguistics, Stroudsburg, PA, USA, 7329–7346. . https://doi.org/10.18653/v1/2021.acl-long.569

[66] Sanjay Subramanian, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner. 2020. Obtaining Faithful Interpretations from Compositional Neural Networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. Association for Computational Linguistics, Stroudsburg, PA, USA, 5594–5608. . https://doi.org/10.18653/v1/2020.acl-main.495

[67] Xiaofei Sun, Diyi Yang, Xiaoya Li, Tianwei Zhang, Yuxian Meng, Han Qiu, Guoyin Wang, Eduard Hovy, and Jiwei Li. 2021. Interpreting Deep Learning Models in Natural Language Processing: A Review. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, November 20, 2021. Association for Computational Linguistics, 20–23. . Retrieved October 18, 2023 from https://arxiv.org/abs/2110.10470

[68] Alona Sydorova, Nina Poerner, and Benjamin Roth. 2019. Interpretable Question

Answering on Knowledge Bases and Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July 28, 2019. Association for Computational Linguistics, Florence, Italy, 4943–4951. . Retrieved October 18, 2023 from http://dx.doi.org/10.18653/v1/P19-1488

[69] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2023. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2023. Association for Computational Linguistics, Stroudsburg, PA, USA, 5644–5659. . https://doi.org/10.18653/v1/2023.acl-long.310

[70] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July 2019. Association for Computational Linguistics, Stroudsburg, PA, USA, 4593–4601. . https://doi.org/10.18653/v1/P19-1452

[71] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. *arXiv:1905.06316* (May 2019). Retrieved October 18, 2023 from https://arxiv.org/abs/1905.06316

[72] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Generating Token-Level Explanations for Natural Language Inference. In *Proceedings of the 2019 Conference of the North*, June 2019. Association for Computational Linguistics, Stroudsburg, PA, USA, 963–969. . https://doi.org/10.18653/v1/N19-1101

[73] Erico Tjoa and Cuntai Guan. 2021. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Trans Neural Netw Learn Syst* 32, 11 (November 2021), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

[74] Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention Interpretability Across NLP Tasks. *arXiv:1909.11218* (September 2019). Retrieved October 18, 2023 from https://arxiv.org/abs/1909.11218

[75] Eric Wallace, Matt Gardner, and Sameer Singh. 2020. Interpreting Predictions of NLP Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, 2020. Association for Computational Linguistics, Stroudsburg, PA, USA, 20–23. . https://doi.org/10.18653/v1/2020.emnlp-tutorials.3

[76] Wenqi Wang, Run Wang, Lina Wang, Zhibo Wang, and Aoshuang Ye. 2021. Towards a Robust Deep Neural Network against Adversarial Texts: A Survey. *IEEE Trans Knowl Data Eng* 35, 3 (October 2021), 3159–3179. https://doi.org/10.1109/TKDE.2021.3117608

[77] Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of Politically Biased Data on Hate Speech Classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, November 2020. Association for Computational Linguistics, Stroudsburg, PA, USA, 54–64. . https://doi.org/10.18653/v1/2020.alw-1.7

[78] Sarah Wiegreffe and Ana Marasović. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS'21)*, February 23, 2021. . Retrieved October 18, 2023 from https://arxiv.org/abs/2102.12060

[79] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, November 13, 2019. Association for Computational Linguistics, 11–20. . Retrieved October 18, 2023 from http://dx.doi.org/10.18653/v1/D19-1002

[80] Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional BERT Contextual Augmentation. In *Proceedings of the International Conference on Computational Science - ICCS 2019*, June 18, 2019. Springer, 84–95. . https://doi.org/10.1007/978-3-030-22747-0_7

[81] Qizhe Xie, Xuezhe Ma, Zihang Dai, and Eduard Hovy. 2017. An Interpretable Knowledge Transfer Model for Knowledge Base Completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2017. Association for Computational Linguistics, Stroudsburg, PA, USA, 950–962. . https://doi.org/10.18653/v1/P17-1088

[82] Yang Yang, Deyu Zhou, Yulan He, and Meng Zhang. 2019. Interpretable Relevant Emotion Ranking with Event-Driven Attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, November 2019. Association for Computational Linguistics, Stroudsburg, PA, USA, 177–187. . https://doi.org/10.18653/v1/D19-1017

[83] Jin Yong Yoo and Yanjun Qi. 2021. Towards Improving Adversarial Training of NLP Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, September 01, 2021. Association for Computational Linguistics, 945–956. . Retrieved October 18, 2023 from http://dx.doi.org/10.18653/v1/2021.findings-

emnlp.81

[84] Muhammad Bilal Zafar, Michele Donini, Dylan Slack, Cedric Archambeau, Sanjiv Das, and Krishnaram Kenthapadi. 2021. On the Lack of Robust Interpretability of Neural Text Classifiers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, August 2021. Association for Computational Linguistics, Stroudsburg, PA, USA, 3730–3740. . https://doi.org/10.18653/v1/2021.findings-acl.327

[85] Lê Nguyễn Nguyên Anh, Nguyễn Quốc Thắng. 2024. Một phương pháp giải thích mô hình SRL dựa trên khái niệm chuyên ngành [Khóa luận cử nhân đại học, Đại học Khoa học Tự nhiên, Đại học Quốc gia Tp.HCM]

Phụ lục

Các bước bổ sung của quá trình tinh chỉnh BioBert:

Finetuning BioBert cho tác vụ SRL bao gồm:

1. Thêm lớp fully connected layer để chuyển đổi đầu ra của BioBert (vector 768 chiều) thành không gian nhãn trong SRL (15 nhãn).
2. Chuẩn bị và mã hóa dữ liệu dưới dạng tensor để phù hợp với yêu cầu đầu vào của mô hình (được trình bày rõ trong phần 5.1).
3. Fine-tune mô hình: Mô hình được huấn luyện bằng cách tối ưu hóa các trọng số thông qua quá trình backward propagation, sử dụng AdamW làm optimizer và CrossEntropyLoss làm hàm mất mát, diễn ra trong 10 epoch.
4. Đánh giá và lưu mô hình sau khi hoàn tất quá trình huấn luyện.